Loughborough University
Institutional Repository

Efficiency and productivity analysis in ten Asian banking industries

This item was submitted to Loughborough University's Institutional Repository by the/an author.

Additional Information:

- A Doctoral Thesis. Submitted in partial fulfillment of the requirements for the award of Doctor of Philosophy of Loughborough University.

Metadata Record: https://dspace.lboro.ac.uk/2134/6110

Publisher: © Zhi Shen

Please cite the published version.

This item was submitted to Loughborough's Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to: http://creativecommons.org/licenses/by-nc-nd/2.5/

Efficiency and Productivity Analysis in Ten Asian Banking Industries

by

Zhi Shen

A Doctoral Thesis
Submitted in partial fulfillment of the requirements
for the award of
Doctor of Philosophy of Loughborough University

December 2009

© by Zhi Shen, 2009

Dedication

To

My true love and wife, Yu Sun

Thank you for your unconditional sacrifice and tolerance.

I owe you my whole life.

Acknowledgement

I am deeply in debt to my leading supervisor, Professor Tom Weyman-Jones, for his tireless guidance, endless support and encouragement through my study. His great vision, profound knowledge and wise suggestions are lifelong treasure to me. Without his help, this work could not be finished.

I would like to express my immense gratitude to Dr. Lawrence Leger for his great effort to improve my English writing and all the academic and social help during the last five years.

I would also like to express my deep gratitude to my second supervisor, Dr. Hailin Liao, for her insightful comments and suggestions to my work.

I am also grateful of the generous financial support from Department of Economics, Loughborough University.

My deepest gratitude is to my family. I want to thank my parents for their unconditional love, life guidance and all the supports. I would also like to thank my parents-in-law for their support and generous understanding for all the troubles I leave with them. I owe my loving thanks to my dear wife Yu Sun and my dear son Canning Shen. I will repay your love, trust and sacrifice with my whole life.

A paper extracted from the fourth chapter of this thesis titled ‘Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions’ (co-authored with Professor Tom Weyman-Jones and Dr. Hailin Liao) is published in Journal of Chinese Economic and Business Studies, May 2009 and in edited book titled ‘China’s Three Decades of Economic Reform’, Aug 2009. I would like to thank participants in the June 2008 fifth North American Productivity Workshop and participants in the April 2008 CEA-UK annual conference for their useful comments on earlier drafts of this paper. Thanks also to the two anonymous referees of the journal and editor Professor Xiaohui Liu for her valuable advice and generous help.

Abstract

Over the last few decades, numerous studies have adopted efficiency and productivity techniques to examine and evaluate the overall performance of banking industries to inform policy effect as well as identify the best practice. The majority of banking efficiency and productivity studies focus on the developed US and European countries. There are only limited studies in the Asian banking industries but no cross-country comparison in major Asian economies. To fill this literature gap, this thesis attempts to measure and compare the cost efficiency and total factor productivity change in ten Asian banking sectors using an unbalanced panel data set consisting of 280 commercial banks over the period of 1998 to 2005. It is widely agreed that cross-country differences play an important part in examining banks' performance in international comparison. They can influence the frontier technology as additional explanatory variables or they can enter inefficiency directly as a measure of determinants or heteroscedasticity. Both cases are considered in the empirical sections of this thesis. In the former case, the empirical results from systematic comparisons of panel data stochastic frontier models with and without incorporating these cross-country heterogeneities suggests that cross-country differences are important sources to explain banks' performances therefore they should not be neglected. The overall cost efficiency in these Asian banking industries is 0.5897 with a decreasing trend, despite positive technical progress and slight economies of scale. The total factor productivity change is measured by using a new cost-based total factor productivity index, an index number counterpart of Bauer's (1990) total differential approach. A five-way decomposition is also provided with the attempt to identify the main contributors to the productivity change. Overall, Asian banking industries have experienced positive but not substantial productivity change from 1998-05. In the latter case, a general model that considers exogenous influences in both inefficiency and random noise error term is constructed and compared against other alternative specifications. The empirical results favour this general model and the overall and country-specific cost efficiency and total factor productivity are then estimated and calculated.

Key words: bank performance, cost efficiency, cross-country heterogeneity, panel data, total factor productivity change, heteroscedasticity

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

The theory of production economics tells the story of optimization. Producers aim to maximize their feasible outputs given the production technology in place and levels of input resources. They also attempt to minimize the production costs under current technology and input prices they face, as well as to maximize profits, given the technology and output and input prices in place. Early econometric approaches to practices of production theory focused on development of flexible functional form of production, cost and profit, sharing the same assumption that producers operate on these functions, apart from the randomly distributed statistical noise, with an effort to learn and exploit the true structure of production technology.

However, numerous empirical evidences suggest that not all producers are always successful in achieving the optimization. Some producers are deemed technically inefficient as they fail to maximize output expansion given the resources they have or minimize the input utilization given the output level they set to produce. Even if they are technically efficient, they may possibly be cost inefficient in the sense of their failure to allocate their inputs in a cost-minimizing manner, given the input prices they face, which may contribute further to the failure to minimize expenditures in output production. Furthermore, even if some producers are cost efficient, not all of them can be profit efficient because of misallocation of outputs in a revenue-maximizing manner, given the output prices, which will result in failure to maximize profits. Consequently, traditional econometric practices of production, cost and profit function associated with symmetrically and randomly distributed statistical noise are no longer appropriate. To study producers' behavior of failure of optimization, a reformulation from production, cost and profit functions to respective frontiers is required so as to allow technically (or cost and profit) inefficient producers to lie beneath the production (or profit) frontiers (for cost efficiency, inefficient producers lie over the cost frontier) and departure from frontier will be treated as inefficiency, apart from randomly distributed statistical noise.

The history of theoretical developments in frontier analysis of producers' performance went back to the pioneer work of Michael Farrell, who was the first to measure productive efficiency. Inspired from Koopmans (1951) and Debreu (1951), Farrell (1957) introduced a method to decompose the overall efficiency into its technical and allocative components. His work influenced the development of data envelopment analysis (DEA for short) by Charnes *et al.* (1978) and stochastic frontier analysis (SFA for short) by Aigner *et al.* (1977) and Meeusen and van den Broeck (1977). DEA and SFA are by now well-established and widely used non-parametric and parametric efficiency measurement techniques in the literature of performance evaluation.

Due to the rapid change in banking operation and regulatory structure all over the world, the banking sector has become one of the largely studied fields in efficiency and productivity literature in the last two decades. Events like large waves of inter-country and cross border mergers and acquisition in the US and European Union, along with banking reforms such as deregulation, privatization process in developing countries, trigger economists' interests to evaluate the performance of financial institutions on purpose to assess the effectiveness of government policy, such as deregulation and privatization on efficiency, to address research issues such as effects of mergers, non-performing loan, market structure, employment of different methodology on efficiency, and to improve managerial performance by identifying the most efficient producers and its driven contributors. The emphasis of this area spreads widely from scale and scope economies to the cost and profit efficiency by applying the above non-parametric and parametric frontier methods such as DEA and SFA.

However, no matter whether DEA or SFA is adopted, most banking efficiency and productivity studies are applied to developed countries such as the US and European Union. Compared to the large volume of the US and European studies, the number of Asian banking efficiency and productivity studies is really small. In the wake of the economic and financial reform, Asian banking industries have experienced dramatic change in operational environment and regulatory structure. Of the great inspiration is to examine the banking performance and effects of deregulation and privatization on efficiency in Asian countries. Moreover, along with the progress of economic and financial integration in Asian economies, it is also interested to see how these banking industries perform against each other. However, so far, there are no cross-country efficiency and productivity studies in major Asian economies. Therefore, this literature gap brings the intuitive motivation to this thesis.

1.2 Thesis overview and aims of this research

This thesis aims to fill the above literature gap by examining and comparing cost efficiency level and measuring total factor productivity change and its decomposition in ten Asian banking industries from 1998 to 2005.

Of the significance of cross-country studies is its usefulness to provide valuable information on the competitiveness of banks in different countries, a very important issue to address in the harmonized European markets of banking services and more globalized financial markets nowadays (Berger and Humphrey, 1997). However, cross-country studies may also be difficult to explain due to differences in market structure and economic and regulatory environment faced by banks and relative banking services provided. Such differences have not been specified in early cross-country studies in which a common frontier technology was estimated. Dietsch and Lozano-Vivas (2000) was among the first few to challenge this common frontier technology presumed in cross-country studies by including seven environmental variables that capture three categories of cross-country differences as additional explanatory variables in the model. They found that without incorporation of cross-country differences, inefficiency scores were overestimated since it captured variations between countries that should be explained as cross-country differences but absorbed as part of the inefficiency due to use of common frontier technology. However, existing banking efficiency studies that control cross-country differences used the cross-sectional stochastic frontier framework and it suffers serious shortcomings. As pointed out in Schmidt and Sickles (1984), these shortcomings can be fixed by using panel data stochastic frontier approaches. Therefore, I intend to examine cross-country exogenous influences in production technology in the context of panel data stochastic frontier models and to see whether Asian banking industries are operating on a common production technology.

In the context of panel data stochastic frontier approaches, traditional time-invariant and time-varying fixed-effects and random-effects models have difficulties in handling the time-invariant heterogeneities. In traditional panel data models, unobservable time-invariant heterogeneities will be pushed into the inefficiency term, resulting in underestimation of the overall efficiency level. Greene's (2004) was perhaps the first to address this problem by introducing his 'true' stochastic frontier models that use firm-specific intercepts to capture unobservable time-invariant heterogeneities. However, he suggested that use of 'true' stochastic frontier models might overcompensate for time-invariant heterogeneities because persistent inefficiency will be treated as time-invariant heterogeneities as well. Consequently efficiency will be overestimated. It is my intention to provide further empirical evidence on treatments of time-invariant heterogeneities by systematically comparing the traditional and 'true' panel data stochastic frontier models in the application to Asian banking data. Using the preferred panel data stochastic frontier approach, competitiveness of commercial banks in different countries is examined by measuring and comparing cost efficiency in these Asian banking industries. Research questions such as trend of efficiency change, economies of scale, appropriate choice of output specification, choice of functional form are also discussed.

The literature on banking productivity measurement has been dominated by studies using non-parametric Malmquist productivity index. The only study adopted cost based parametric approach to measure the total factor productivity (TFP for short) change and its attributes was Bauer (1990). However, since Bauer's total differential approach was constructed under continuous time it cannot allow productivity change in yearly basis. It is desirable to develop a new index number counterpart to Bauer (1990) but allow to measure of the TFP change in discrete time. Using short run cost function and Diewert's (1976) Quadratic Identity Lemma, a Malmquist type productivity index is constructed and further decomposed to five contributes, termed as cost efficiency change, technical change, change of scale effects, and two allocative efficiency change components, output mix allocative efficiency change and input mix allocative efficiency change, inclusion of which in the measurement of TFP change has not been seen in previous studies.

In applications of measuring efficiency of international airlines, Coelli *et al.* (1999) pointed out that exogenous environmental factors may not only influence the production frontier technology but also have direct impacts on inefficiency term as determinants of inefficiency (also see Kumbhakar *et al.*, 1991, Huang and Liu, 1994, and Battese and Coelli, 1995). In addition, exogenous influences on inefficiency may also come from affecting variances of inefficiency as a control of heteroscedasticity (see Reifschneider and Stevenson, 1991, Caudill and Ford, 1993, Caudill *et al.*, 1995 and Hadri, 1999). Moreover, Hadri (1999) and Hadri *et al.* (2003a, 2003b) also considered possible exogenous environmental impact on the symmetrically distributed random noise term. This thesis attempts to examine these three potential exogenous environmental influences on composed error term by estimating a general stochastic cost frontier model that incorporates cross-country environmental influences on both mean and variances of cost inefficiency and variances of random noise term. Models with alternative specifications are tested against this general model. Cost efficiency and total factor productivity change are estimated and calculated from the best fitted model.

1.3 Contribution to knowledge

This thesis has six main contributions. First of all, this thesis fills the literature gap in banking efficiency and productivity studies that have mainly focused on banking industries in the US and European countries by examining and comparing cost efficiency and total factor productivity change and its determinants in ten Asian banking industries. Although there are banking efficiency and productivity studies in individual Asian country and one cross-country comparison in four South Asian countries (Perera *et al.*, 2007), no international comparison has been carried out between major Asian economies.

Second, unlike most banking efficiency studies using cross-sectional stochastic frontier approach, this thesis adopts panel data stochastic frontier approaches and systematically compares and analyzes the pros and cons of time-invariant and time-varying fixed- and random-effects models.

Third, this thesis is among the first to examine the exogenous cross-country heterogeneous impacts on not only the frontier technologies but also inefficiency and random noise terms. The exogenous impacts on frontier technologies are modeled as adding environmental variables that reflects cross-country heterogeneities as additional explanatory variables in the cost frontier, while exogenous influences on the composed error term are modeled by allowing these cross-country environmental variables to enter the mean of inefficiency term as determinants or/and to enter the variance of inefficiency and random error terms as a measure of controlling heteroscedasticity.

Fourth, this thesis uses a new approach to measure the total factor productivity change and identify its determinants. Other than using a non-parametric Malmquist productivity index, this thesis has developed a parametric cost based Malmquist-type productivity index, inspired by Bauer (1990) in which the author constructed a total differential approach to measure the total factor productivity change. A new five-way decomposition of the total factor productivity change is created and the importance of considering the impact of allocative efficiency change in the context of both input and output mix is also highlighted.

Fifth, this thesis also contributes to the literature by providing tests on the theoretical properties associated with cost function. In theory, the monotonicity condition suggests that cost function should be non-decreasing in outputs and input prices, while the concavity condition requires cost function to be concave in input prices. Most banking efficiency studies ignore these tests and some of the estimated parameters fail to satisfy these essential monotonicity and concavity conditions. In these circumstances the creditability of estimated efficiency results is doubtful and corresponding policy implications and suggestions may be misleading.

Last but not least, this thesis provides the derivation of probability density function of inefficiency conditional on composed error term. This was provided in the original paper of Aigner *et al.* (1977) without any proof. Since it is taken as granted in most efficiency studies, I feel that providing the derivation can be helpful to understand the core legacy of the stochastic frontier approach and may be useful for researchers who would like to develop and incorporate modified specifications to the existing frontier models and to write ones' own program for estimation without relying on the existing software such as LIMDEP and STATA, which may be limited in applying certain new model specifications.

1.4 Structure of this thesis

The whole thesis is organized as follows.

Chapter 2 provides detailed literature review in the context of both empirical banking efficiency studies and theoretical methodologies to measure efficiency and productivity changes. Section 2.1 surveys banking efficiency studies using the parametric stochastic frontier approaches. This general survey is categorized into six respective themes – applications to different countries, choices of variables used in estimation, choices of functional form, economies of scale and scope, efficiency estimates and determinants of efficiencies. Section 2.2 introduces the origin and recent developments of frontier efficiency measurement methodology with a main focus on the parametric stochastic frontier approach. The cross-sectional stochastic frontier framework that is usually adopted in banking efficiency studies is introduced, followed with discussions of latest development of panel data stochastic frontier framework. Methodological review of productivity measurement is presented in section 2.3. Discussions are carried out in the context of non-frontier and frontier methods to measure productivity change. Both methodologies are further categorized into non-parametric and parametric methods.

Chapter 3 explains the main purpose of this thesis and provides a detailed description of sample data set. My data consists of 280 commercial banks from ten Asian banking industries from 1998 to 2005. Choices and definitions of dependent and independent variables are clearly demonstrated as well as the environmental variables that capture cross-country differences.

Chapter 4 studies cost efficiencies of ten Asian banking industries with incorporation of the impact of cross-country heterogeneous factors on the frontier technologies. Various panel data stochastic frontier models such as time-invariant and time-varying fixed-effects and random-effects models are modeled and compared. The assumption of whether banks from different countries share the same frontier technology are being tested with specifications of including or excluding cross-country environmental variables in the models. Overall and country specific cost efficiency scores are estimated based on parameter estimates from the best fitted model. This chapter ends with discussions on two research questions that induce a long-standing debate that reaches no general consensus. The first issue concerns about the appropriate choice of input and output specifications while the other relates to the appropriate choice of flexible functional form.

Based on the parameter and efficiency estimates from preferred model in Chapter 4, Chapter 5 measures the total factor productivity change and its decomposition on the purpose to find the potential sources driving the productivity progress (or deterioration). In order to develop the new approach to measure the TFP change, Bauer's (1990) total differential approach on TFP decomposition is derived in the context of multiple outputs. Based on that, a Malmquist type productivity index, the new index number approach, counterpart to total differential approach, which allows comparisons of TFP change year by year, is constructed with a five-way decomposition generated. Overall and country specific TFP change is calculated and its main driven sources are identified.

Chapter 6 examines the cross-country environmental influences on the inefficiency term as well as the random noise term. This chapter starts from a detailed review of existing literature addressing this issue. A general panel data stochastic frontier model comprising specifications of these exogenous influences on the mean and variances of the inefficiency term and variances of the random noise term is constructed. Six alternative specifications are tested against this general model using the likelihood ratio test and the Wald test. Overall and country-specific cost efficiency and TFP change are estimated and calculated using parameter estimates from the best fitted model. Cost efficiency trend over the sample period is discussed, as well as the main driven attributes to the TFP change.

Chapter 7 concludes the thesis and directions for future researches are also provided.

Chapter 2 Literature Review

2.1 Literature review of banking efficiency studies

2.1.1 Introduction

Ever since the pioneer seminar paper of Farrell (1957) that introduced a frontier method to measure firms' economic efficiency, there have been numerous efficiency studies in financial institutions and mostly banks. Some studies tried to inform government policies by assessing the effect of deregulation, loan quality and risk factor, market structure and mergers and acquisitions on firms' performance. Some focused on improving the frontier methodology to obtain consistent and more accurate efficiency estimates while others concentrated on improving the managerial performance by identifying the "best" and "worst" practice.

The first extensive survey of efficiency studies in financial institutions was provided by

Berger *et al.* (1993), in which the authors mainly reviewed scale and scope efficiencies and X-efficiency in banking, along with discussions of mergers, efficiencies in governmental financial institutions and insurance companies, and determinants of efficiency of financial institution. It was clear that up to that date, the banking efficiency literature was dominated by studies of scale and scope efficiencies rather than X-efficiencies. The second and most exhaustive empirical survey was provided by Berger and Humphrey (1997). Their paper surveyed 130 studies that apply the frontier efficiency analysis to financial institutions in 21 countries with an attempt to reach a consensus view. Various frontier efficiency methods have been critically reviewed with suggestions of ways that these methods might be improved to bring more accurate and useful efficiency estimates. Critical discussions were also provided in terms of the following three categories: policy issues such as effect of deregulation, non-performing loan, market structure and mergers and acquisitions; research issues such as confidence interval, comparisons of different efficiency techniques and assumptions, international comparison, X-efficiency; and research issues of improving managerial performance. Their survey concluded by outlining potential directions for future research such as improvement of frontier efficiency techniques and research issues of providing more evidence in cross-country comparison and determinants of efficiency estimates.

Despite no consensus on the “best” frontier method, I favour the stochastic frontier approach because of its virtue of allowing random noises that are outside control of firms and comprising measurement error, specification error and sampling error. Although stochastic frontier approach is criticized by imposing a strict functional form that presuppose the shape of an unknown frontier, I think the risk of misspecifying the true frontier is less than the risk of ignoring it, not even saying that the risk of misspecification could be controllable by running statistical and econometric test on the model and the use of explanatory variables and test on theoretical properties of the presumed functional form. Therefore, unlike the above surveys, my own literature survey of banking efficiency studies focuses on those studies using stochastic frontier approach and on practical issues that would shed light on my own research interest and motivation.

2.1.2 Survey of banking efficiency studies using stochastic frontier approach

2.1.2.1 Applications to different countries

There has been rapid change in the financial industry around the world over the last two decades. In the US, the banking industry has been through dramatic waves of consolidations in the 1990s and recently was the centre of financial crisis initiated from subprime loan crisis. In the Europe Union, there have been a large number of consolidations of banks within countries and across borders in Western Europe and considerable deregulation and privatization in financial system of Eastern Europe, as part of the progress of EU integration. In South and East Asia, countries are reconstructing financial systems and regulatory structure after suffering the devastating damage of the 1997 Asian financial crisis. In Latin American and the Caribbean, an intense wave of foreign bank entry and cross-border mergers and acquisitions has been witnessed since the financial crisis during the initial liberalization period.

To keep pace with these financial revolutions and structural change, a large number of efficiency studies appear in the literature with an attempt to assess the effect of government policy such as deregulation, privatization and mergers on banks' overall performance and provide further policy implications. The majority of studies in my survey are applied into the US and developed European countries for both country-specific level and international comparisons (see Figure 2.1 and 2.2). There are ten Asian banking efficiency studies with two studying Japanese banks, three for Chinese banking sector and one for Korea, Taiwan Province of China, Hong Kong SAR, India respectively. Only one cross-country comparison is found in four South Asian countries. An apparent literature gap in efficiency studies using stochastic frontier approach lies in the lack of international comparisons of major Asian banking industries, which may partly subject to the lack of quality data in these countries and regions.

2.1.2.2 Variables

No matter what kind of frontier analysis technique one is using and no matter what kind of models or functions one is adopting, obtaining accurate and reliable efficiency estimates heavily relies on the quality of data and variables used in regression. Some efficiency studies may share the same manner in using explanatory variables but some may not. Regarding to the absence of coincidence in variables involved, it is necessary to review the literature and critically summarize the existing manner of variables in use.

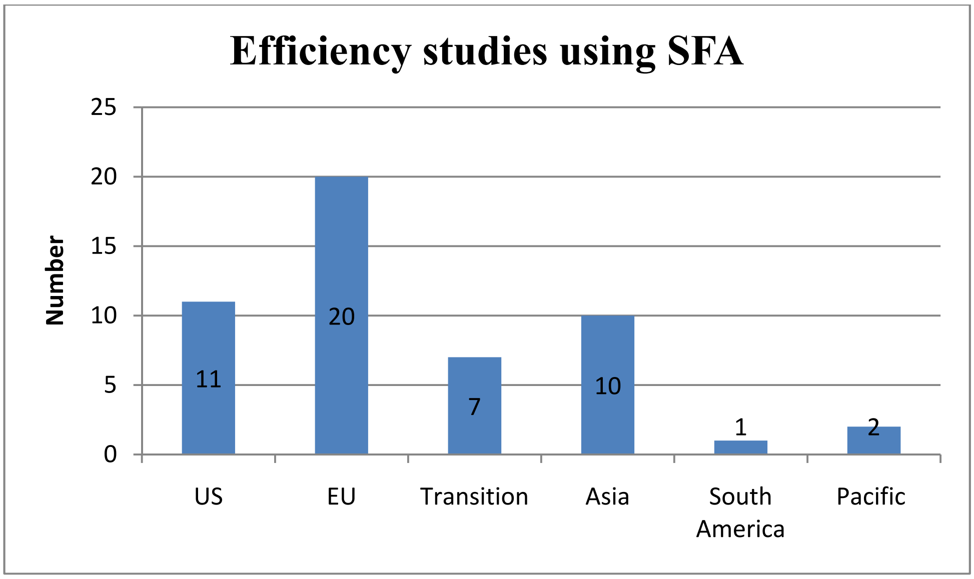

Figure 2.1: Distribution of studies using SFA

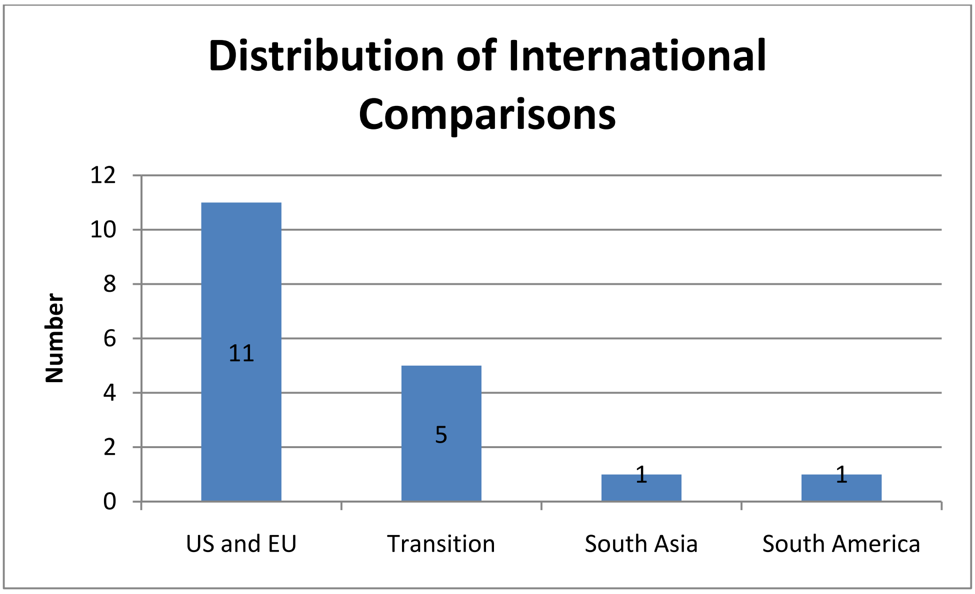

Figure 2.2: Distribution of international comparisons

2.1.2.2.1 Input and output specifications

Before any model of production and cost frontier can be developed, the input and outputs of the banking firm should be clearly defined. No general consensus exists concerned about the precise definition of what bank produce or how one can measure banks' products. Financial firms, however, provide services rather than readily identifiable physical products, and it is not clear how to measure service outputs. The definition and measurement of bank output and costs vary between empirical studies. With respect to the role of deposits, I find there are four kinds of output and input specifications which are commonly used in the modern literature of banking efficiency analysis. They are intermediation approach, production approach, dual approach and value added approach.

Twenty six out of forty eight efficiency studies in my survey, presented in Table 2.1, adopt the **intermediation approach** introduced by Sealey and Lindley (1977) (Altunbaş *et al.*, 2000; Altunbaş *et al.*, 2001; Berger and DeYoung, 1997; Berger *et al.*, 1997; Berger and Mester, 1997; Carbo *et al.*, 2002; Casu and Girardone, 2004; Casu and Girardone, 2002; Chang *et al.*, 1998; Christopoulos *et al.*, 2002; Christopoulos and Tsianos, 2001; Cuesta and Orea, 2002; Girardone *et al.*, 2004; Huang, 2000; Kaparaikis *et al.*, 1994; Kumbhakar and Wang, 2007; Kwan, 2002; Kwan and Eisenbeis, 1996; Lang and Welzel, 1996; Mertens and Urga, 2001; Mester, 1996; Perera *et al.*, 2007; Sturm and Williams, 2005; Vander Vennet, 2002 and Weill, 2004, 2003). They investigated what product banks are truly producing and suggested that there are two kind of production process involving technical and economic production. The technical process of production is a transformation process in which banks borrow funds from surplus spending units and lend those funds to the deficit spending units, known as the financial intermediation. The economic process requires banks to create a product which is more highly valued than the original input elements. Services provided for the depositors cannot create such high value since they are more appropriately associated with the acquisition of economic inputs which require banks to incur physical costs without yielding any direct revenue. In other words, as partial payments for the use of funds from depositors, banks produce, at positive costs of capital and labour, services such as safekeeping, check clearing, bookkeeping, etc. to the depositors. Therefore,

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches

PANEL A: COST EFFICIENCY

Studies	**Applied Countries**	**Functional Form** [a]	**Dependent variable**	**Independent Variables**			**Data**	**Average efficiency estimates** [b]
				Outputs	**Input prices**	**Other**		
Allen and Rai (1996)	15 EU countries	TL	Total cost	Y1-Total loans Y2-Investment securities	W1-Price of labour W2-Price of funds W3-Price of capital		15 countries between 1988 to1992	0.85 (Universal), 0.79 (Separated)
Altunbaş *et al.* (2001)	EU	FF	Total cost /operating and financial cost	**Intermediation** Y1-Total loans Y2-Total securities Y3-Total off-balance sheet items	W1-Price of labour W2-Price of funds W3-Price of physical capital	Equity capital is included to control for difference in bank's risk reference.	EU banks between 1989 and 1997 with 4104 observations	0.75 - 0.8 across different asset sizes
Altunbaş *et al.* (2000)	Japan	FF	Total cost	**Intermediation** Y1-Total loans Y2-Total securities Y3-Total off-balance sheet items	W1-Price of labour W2-Price of funds W3-Price of physical capital	Equity capital, NPL ratio to control for the risk and quality.	139 banks for each year from 1993 to 1995 and 136 in 1996	Inefficiency level range between 5% and 7% with no discernible trend across size classes.
Berger and DeYoung (1997)	US	FF	Operating cost (only include non interest expense)	**Intermediation** Y1-Commercial loans Y2-Consumer loans Y3-Real estate loans Y4-Transaction deposits Y5-Fee-based income	W1-Price of labour W2-Price of physical capital		U.S. commercial banks from 1985 to 1994	0.92
Berger *et al.* (2009)	China	TL	Total cost	**Dual** Y1-Total loans Y2-Total deposits Y3-Liquid assets Y4-Other earning assets	W1-Interest expense to total assets W2-Non-interest expenses to fixed assets	Z1- total earning asset	37 commercial banks from 1994-2003	0.74

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variables			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Berger *et al.* (1997) use DFA	US	FF	Total cost	**Intermediation** Y1-Value of consumer transaction accounts Y2-Value of consumer nontransaction accounts Y3-Value of business transaction accounts Y4-Value of business nontransaction accounts	W1-Average wage rate W2-Average rental rate on capital		832 offices of an anonymous large US commercial bank for 1989, 1990 and 1991	0.9-0.95
			Operating cost	**Production** Y1-Number of deposit accounts Y2-Number of debits Y3-Number of credits Y4-Number of accounts opened Y5-Number of accounts closed Y6-Number of loans originated				0.75-0.8
Berger and Mester (1997)	USA	TL, FF	Total cost	**Intermediation** Y1-Consumer loans Y2-Business loans (all other loans) Y3-Securities (all non-financial assets)	W1-Price of purchasing funds (liabilities except core deposits) W2-Price of core deposits W3-Price of labour	Z1-off-balance-sheet Z2- Physical capital Z3- Financial equity capital	Almost 6000 U.S. commercial banks from 1990-1995	0.87 (FF,DFA) 0.86 (TL,DFA)
Bonin *et al.* (2005a)	11 transition countries	TL	Total cost	**Value-added** Y1-Total deposits Y2-Total loans Y3-Total liquid assets Y4-Investments	W1-Price of capital W2-Price of funds	Dummies of year and country effects	225 banks from 1996 to 2000	0.78

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Bonin *et al.* (2005b)	6 transition countries	TL	Total cost	**Value-added** Y1-Total deposits Y2-Total loans Y3-Total liquid assets Y4-Investments	W1-Price of capital W2-Price of funds	Dummies of year and country effects	67 banks from 1994 to 2002	0.79
Casu and Girardone (2004)	5 EU countries	FF, TL	Total cost	**Intermediation** Y1-Total loans Y2-Other earning assets	W1-Price of labour W2-Price of funds W3-Price of fixed assets		banks covering 1993 to 1997	0.86 (FF), 0.87 (TL)
Carbo *et al.* (2002)	12 EU countries	FF	Total cost	**Intermediation** Y1-Loans Y2-Securities Y3-Off-balance sheet activities	W1-Price of labour W2-Price of purchased funds W3-Price of physical capital		saving banks from 1989 to 1996	0.78
Carvallo and Kasman (2005)	Latin American and Caribbean countries	TL	Total cost	**Dual** Y1-Loans Y2-Deposits Y3-Other earning assets (investment securities)	W1-Price of labour W2-Price of purchased funds W3-Price of physical capital	Country-specific environmental variable	481 banks in 16 Latin American and Caribbean countries over 1995-1999	0.78
Cavallo and Rossi (2002)	6 EU countries	TL	Total cost	**Dual** Y1-Loans Y2-Deposits Y3-Investments	W1-Price of labour W2-Price of deposits W3-Price of fixed assets	Size dummies, organizational dummies and bank balance indicators	banks covering 1992 to 1997	
Chang *et al.* (1998)	USA	TL	Total cost	**Intermediation** Y1-All money market assets Y2-Commercial and industry loans Y3-Other loans Y4-Other bank output	W1-Price of labour W2-Price of physical capital W3-Price of funds		1472 foreign-owned and US owned multinational banks from 1984-1989	0.73 (Foreign owned multinational banks) 0.79 (US owned multinational banks)

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Christopoulos and Tsionas (2001)	Greece	CD	Total cost	**Intermediation** Y1-Loans Y2-Investments Y3-Liquid assets	W1-Price of funds W2-Price of labour W3-Price of capital		19 banks from 1993 to 1998	0.68
Christopoulos *et al.* (2002)	Greece	TL	Total cost	**Intermediation** Y1-Loans Y2-Investments Y3-Liquid assets	W1-Price of funds W2-Price of labour W3-Price of capital		19 banks from 1993 to 1998	0.6-0.95
Dietsch and Lozano-Vivas (2000)	France and Spain	TL	Total cost	**Value-added** **Y1-Deposits** **Y2-Loans** **Y3-Other earning assets**	W1-Price of labour W2-Price of physical capital W3-Price of financial factor	7 environmental variables (EV for short)	223 French banks and 101 Spanish banks from 1988 to 1992	Without EV 0.58 Fr., 0.09 Sp. With EV 0.88 Fr., 0.75 Sp.
Ferrier and Lovell (1990)	USA	TL	Total cost	**Production** Y1-Number of demand deposit accounts Y2-Number of time deposit accounts Y3-Number of real estate loans Y4-Number of installment loans Y5-Number of commercial loans	W1-Price of labour W2-Occupancy costs and expenditure on furniture and equipment over level of deposits W3-Expenditure on materials over level of deposits	12 EV	575 institutions which participated in the FCA programme in 1984.	0.74 (SFA), 0.78 (DEA)
Fries and Taci (2005)	15 transition countries	TL	Total cost	**Value-added** Y1-Loans Y2-Deposits	W1-Price of labour W2-Price of physical capital	8 country-level EV	289 banks from 1994 to 2001	Without EV: 0.63 With EV: 0.71

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Fu and Heffernan (2007)	China	TL	Total cost	**Dual** Y1-Total deposits Y2-Total loans Y3-Investments Y4-Non-interest Income	W1-Price of funds W2-Price of fixed assets W3-Price of employees	z- total asset	14 commercial banks from 1985-2002	0.41 (H), 0.52 (E), 0.46 (T)
Girardone *et al.* (2004)	Italy	FF	Total cost	**Intermediation** Y1-Loans Y2-Other earning assets	W1-Price of funds W2-Price of labour W3-Price of fixed assets	Financia capital and asset quality	545 banks from 1993 to 1995	94.8, 78.5
Hao *et al.* (2001)	South Korea	TL	Total cost	**Dual** Y1-Total loans and securities Y2-Demand deposits Y3-Fee income	W1-Wage rate W2-Interest for borrowed funds W3-Price of physical capital	equity capital adjust for increased costs of funds due to financial risk	19 private Korean banks from 1985 to 1995	0.89
Hasan and Marton (2003)	Hungary	TL	Total cost	**Dual** Y1-Total loans Y2-Total investments Y3-Non-interest or fee-related income Y4-Total interest of borrowed fund	W1-Price of fund W2-Price of labour and related expense	equity capital and loan loss provision as a control for risk preferences, loan quality	commercial banks during 1993 to 1998 with 193 bank observations	0.71 (all sample) 0.67 (all domestic) 0.74 (Foreign investment)
Kaparakis *et al.* (1994)	USA	TL	Total cost	**Intermediation** Y1-Individual loans Y2-Real estate loans Y3-Commercial loans Y4-Trading accounts securities, assets	W1-Price of deposits W2-Price of funds W3-Price of labour W4-Price of capital		5548 US banks in 1986	0.9

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Kasman and Yildirim (2006)	8 Central and Eastern European countries	FF	Total cost	**Dual** Y1-Loans Y2-Total deposits Y3-Other earning assets	W1-Price of funds W2-Price of labour W3-Price of capital	11 country-level EV	190 banks from 1995 to 2002	0.81
Kraft and Tirtiroglu (1998)	Crotia	TL	Total cost	**Dual** Y1-Loans Y2-Deposits	W1-Price of funds W2-Price of labour W3-Price of capital		43 banks in both 1994 and 1995	0.55-0.88
Kwan and Eisenbeis (1996)	USA	TL	Total cost	**Intermediation** Y1-Investment securities Y2-Real estate loans Y3-Commercial and industrial loans Y4-Consumer loans Y5-Off-balance sheet commitment	W1-Price of capital W2-Price of funds W3-Price of labour		254 bank holding companies, of which 174 had complete time-series data from 1986 through 1991	0.8-0.9
Lang and Welzel (1996)	Germany	TL	Total cost	**Intermediation** Y1-Short-term loans to non-banks Y2-Long-term loans to non-banks Y3-Loans to banks Y4-Other investments Y5-Fees and commissions Y6-Commodities sale revenue	W1-Price of labour W2-Price of capital W3-Price of deposits	Branch variable; *merger* dummy interactive with input prices.	757 German cooperative banks between 1989 and 1992	0.69 (FE), 0.54 (RE)
Maudos *et al.* (2002) use DFA, FE and RE	10 EU countries	TL	Total cost	**Dual** Y1-Loans Y2-Other earning assets Y3-Deposits	W1-Price of deposits W2-Price of labour W3-Price of physical capital		832 banks over 1993 to 1996	0.83 (DFA), 0.77 (RE), 0.84 (FE)

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Maggi and Rossi (2003) use DFA	EU, US	TL, FF, Box-Cox	Total cost	**Modified Production for EU** Y1-Deposits Y2-Loans Y3-Services	W1-Price of labour W2-Price of capital W3-Price of deposits		338 European commercial banks and 279 US commercial banks from 1995 to 1998	EU: 0.64 (FF), 0.68 (TL) US: 0.62 (FF), 0.63 (TL)
				Value-added Y1-Deposits Y2-Loans Y3-Services	W1-Price of labour W2-Price of capital			
Maudos and Pastor (2001)	US, Europe, Japan	TL	Total cost	**Dual** Y1-Loans Y2-Other earning assets Y3-Deposits	W1-Price of deposits W2-Price of labour W3-Price of physical capital		Banks over 1984 to 1995	around 0.85 (EU), 0.97 (US), 0.95 (Japan)
Merton and Urga (2001) also use TFA	Ukraine	TL		**Intermediation** Y1-Interbank loans Y2-Client loans Y3-Investments in securities and other investments	W1-Price of funds W2-Price of labour	capital, fixed assets, non-performing ratio	79 banks for 1998	0.67 (SFA), 0.81 (TFA)
Mester (1996)	US	TL	Total cost	**Intermediation** Y1-Real estate loans Y2-Commercial and industrial loans, etc. Y3-Loans to individuals	W1-Price of labour W2-Price of physical capital W3-Price of borrowed money	Average of nonperforming loans and average volume of equity capital in 1992	214 banks from 1991-1992	0.91-0.94
Pastor and Serrano (2005) use DFA	7 EU countries	TL	Total cost	**Dual** Y1-Loans Y2-Other earning assets Y3-Deposits	W1-Price of deposits W2-Price of labour W3-Price of physical capital	13 country-level EV	596 banks from 1993 to 1997	0.89, 0.88 (risk adjusted)

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Perera *et al.* (2007)	4 South Asian countries	TL	Total cost	**Intermediation** Y1-Net total loans Y2-Other earning assets	W1-Price of funds W2-Price of labour W3-Price of capital	8 country-level EV	111 commercial banks from 1997 to 2004	0.89
Resti (1997) also use DEA	Italy	TL	Operating cost	**Production** Y1-Loans Y2-Deposits Y3-Non-interest income	W1-Price of labour W2-Price of capital		270 banks from 1988 to 1992	0.7
Sensarma (2006)	India	TL	Total operating cost	**Value-added** Y1-Fixed deposits Y2-Saving deposits Y3-Current deposit Y4-Investments Y5-Loans Y6-Number of Branches	W1-Price of labour W2-Price of capital	Time, Deregulation dummy, Size, Ownership dummy, Labor and Capital	83 Indian banks from 1986 to 2000	Without group dummies 0.92 (S), 0.94 (P), 0.44 (F), 0.63 (NP) With group dummies 0.91 (S), 0.80 (P), 0.26 (F), 0.10 (NP)
Vander Vennet (2002)	EU	TL, FF	Total cost	**Intermediation** Y1-Total loans Y2-Total securities Y3-Total non-interest income	W1-Price of labour W2-Price of physical capital W3-Price of deposits	Equity capital included	2375 EU banks from seventeen countries for the years 1995 and 1996	0.7 for traditional intermediation output; 0.8 for output mix of traditional and nontraditional activities
Weill (2004) also use DEA, DFA	EU	FF	Total cost	**Intermediation** Y1-Loans Y2-Investment assets	W1-Price of labour W2-Price of physical capital W3-Price of borrowed funds	Equity capital included	688 banks from 1992-1998	0.71 Fr., 0.83 Ger., 0.84 It., 0.78 Sp., 0.66 Sw.
Weill (2003)	Poland and Czech Republic	TL	Total cost	**Intermediation** Y1-Loans Y2-Investment assets	W1-Price of labour W2-Price of physical capital W3-Price of borrowed funds	Equity capital and country dummy	31 Polish and 16 Czech banks for 1997	0.70 (Foreign owned), 0.62 (Domestic owned)

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

PANEL B: PROFIT EFFICIENCY

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable			Data	Average efficiency estimates [b]
				Output	Input prices	Other		
Berger and Mester (1997)	USA	TL, FF	Profit	**Intermediation** Y1-Consumer loans Y2-Business loans (all other loans) Y3-Securities (all non-financial assets)	W1-Price of purchasing funds (liabilities except core deposits) W2-Price of core deposits W3-Price of labour	Z1- off-balance sheet activities Z2- Physical capital Z3- Financial equity capital	Almost 6000 U.S. commercial banks from 1990-1995	Standard profit: 0.55(FF,DFA), 0.54 (TL, DFA) Alternative profit: 0.46 (FF, DFA), 0.45 (TL,DFA)
Bonin *et al.* (2005a)	11 transition countries	TL	Net Profit	**Value-added** Y1-Total deposits Y2-Total loans Y3-Total liquid assets Y4-Investments	W1-Price of capital W2-Price of funds	Dummies of year and country effects	225 banks from 1996 to 2000	0.69
Bonin *et al.* (2005b)	6 transition countries	TL	Net profit	**Value-added** Y1-Total deposits Y2-Total loans Y3-Total liquid assets Y4-Investments	W1-Price of capital W2-Price of funds	Dummies of year and country effects	67 banks from 1994 to 2002	0.45
Casu and Girardone (2004)	5 EU countries	FF, TL	Gross profit	**Intermediation** Y1-Total loans Y2-Other earning assets	W1-Price of labour W2-Price of funds W3-Price of fixed assets		banks covering 1993 to 1997	0.76 (FF), 0.75 (TL)
Huang (2000)	Taiwan	TL	Profit	**Intermediation** Y1-Investments Y2-Loans	W1-Price of labour W2-Price of deposits and borrowed money	Non-performing loan and financial capital	22 Taiwan's bank from 1981 to 1995	
Kasman and Yildirim (2006)	8 Central and Eastern European countries	FF	Profit	**Dual** Y1-Loans Y2-Total deposits Y3-Other earning assets	W1-Price of funds W2-Price of labour W3-Price of capital	11 country-level EV	190 banks from 1995 to 2002	0.63

Table 2.1: Literature survey of banking efficiency studies using stochastic frontier approaches (continued)

Studies	Applied Countries	Functional Form [a]	Dependent variable	Independent Variable: Output	Independent Variable: Input prices	Independent Variable: Other	Data	Average efficiency estimates [b]
Lozano Vivas (1997) use TFA	Spain	TL	Profit	**Dual** Y1-Loans Y2-Deposits Y3-Interbank loans	W1-Price of funds W2-Price of labour W3-Price of capital	Fixed assets, Number of branches and other dummies	54 saving banks from 1986 to 1991	0.72 (Non standard), 0.58 (standard)
Maudos *et al.* (2002) use DFA, FE and RE	10 EU countries	TL	Operating profit	**Dual** Y1-Loans Y2-Other earning assets Y3-Deposits	W1-Price of deposits W2-Price of labour W3-Price of physical capital		832 banks over 1993 to 1996	0.45 (DFA), 0.52 (RE), 0.22 (FE)
Maudos and Pastor (2001)	US, Europe, Japan	TL	Net income and Profit before tax	**Dual** Y1-Loans Y2-Other earning assets Y3-Deposits	W1-Price of deposits W2-Price of labour W3-Price of physical capital		Banks over 1984 to 1995	For net income/Profit before tax 0.7/0.45 (EU), 0.75/0.55 (US), 0.5/0.5 (Japan)
Merton and Urga (2001) also use TFA	Ukraine	TL	Profit	**Intermediation** Y1-Interbank loans Y2-Client loans Y3-Investments	W1-Price of funds W2-Price of labour	capital, fixed assets, non-performing ratio	79 banks for 1998	0.72 (SFA), 0.66 (TFA)
Pastor and Serrano (2005) use DFA	7 EU countries	TL	Operating profit	**Dual** Y1-Loans Y2-Other earning assets Y3-Deposits	W1-Price of deposits W2-Price of labour W3-Price of physical capital	13 country-level EV	596 banks from 1993 to 1997	0.52, 0.57 (risk adjusted)
Vander Vennet (2002)	EU	TL, FF	Profit	**Intermediation** Y1-Total loans Y2-Total securities Y3-Total non-interest income	W1-Price of labour W2-Price of physical capital W3-Price of deposits	Equity capital included	2375 EU banks from seventeen countries for 1995 and 1996	0.68

Notes:

a. TL denotes translog function, FF denotes Fourier flexible functional form

b. Fr. stands for France, Ger. stands for Germany, It. stands for Italy, Sp. stands for Spain, Sw. stands for Switzerland. (H) represents half normal distribution, (E) represents exponential distribution and (T) represents truncated normal distribution. (S), (P), (F), (NP) denotes state-owned banks, private banks, foreign banks and new private banks respectively. DFA is the short word for distribution free approach, TFA implies Thick frontier approach and FE and RE stands for fixed- and random-effects model.

banks' production process is a multistage production process involving intermediate outputs, where loanable funds, borrowed from depositors and serviced by the firm with the use of capital, labour and material inputs, are used in the production of earning assets. Therefore, the appropriate concepts of outputs from the view of financial firm's decision making process are the services provided to the debtors of financial institutions. The value of loans and investments is the appropriate measure of bank output under this treatment, while deposits and cost involving in the production process such as capital, labour are measured should be measured as inputs. Consequently, the operating costs and interest expense are measured as the total costs.

Under this treatment, although there are maybe slight differences in the output specification, normally three variables are used as outputs: total loans which include all types of loans like domestic loans, foreign loans and trust a/c loans; other earning assets including trading securities, public bonds, other investments and equity investments and off-balance sheet items. However, the studies of Berger and DeYoung (1997), Berger and Mester (1997), Chang *et al.* (1998), Kaparaikis *et al.* (1994), Kwan (2002), Kwan and Eisenbeis (1996), Lang and Welzel (1996) and Mester (1996) disaggregate total loans into different loan lines with concerns about the output quality. Ideally, the output vector in the production transformation should be measured as quality-adjusted output. That is, one unit of an output included should be one unit of the output of a particular quality. However, in the cost function estimation, typically the unit of output measurement does not hold constant quality. Therefore, disaggregating total loans into different loan lines (e.g., commercial loans, consumer loans, real estate loans, short-term loans and long-term loans, etc.) goes in the right directions that loans in different categories have different risk characteristics. But unfortunately, these studies don't go far enough, since the loans within a particular category can still have different risks. This problem is addressed in Hughes and Mester (1993), Mester (1996), Altunbaş *et al.* (2000 and 2001) by introducing the quality factor in the cost function (see section 2.1.2.2.2). With respect to the input specification, labour, deposits and physical capital are used in the production analysis and price of labour, price of funds and price of physical capital are used in the cost function.

However, **production approach**, which mainly characterized the literature up to the early 1980s, views banks as institutions that primarily produce financial services for the account holders. Banks process transactions and documents for the customers such as loan applications, credit reports, cheques or other payment instruments, and insurance services. Under this approach, number of deposit account holding and loans outstanding or number of transactions performed on each type of product is measured as banks' outputs. In this case, only physical inputs such as labour and capitals are counted as banks' inputs. Interest payments for the use of funds are not considered as an input. Therefore, total costs include all the operating costs but exclude interest expenses. Ferrier and Lovell (1990) adopt this approach and their defined outputs consist of five components: number of demand deposit accounts, number of time deposits accounts, number of real estate loans, number of instalment loans and number of commercial loans. At the meanwhile, input prices used in the cost function are price of labour, occupancy costs and expenditure on the equipment over level of deposits and expenditure on materials over level of deposits. Similar output and input prices specifications were also used in Berger *et al.* (1997) and Resti (1997), whereas a modified production approach was adopted by Maggi and Rossi (2003) that used the same output specification but included price of deposits as additional input price variable.

Both approaches have some advantages. The intermediation approach has the advantage of being more inclusive and capturing the essence of a financial intermediary. Also it is superior of evaluating branch profitability-total operating plus interest costs per dollar of deposits is a good indicator of profitability, since revenues per dollar of deposits are virtually identical across branches. The production approach is advantageous if the branch is thought of as a producer of depositor services of the bank, which then makes decisions on how to intermediate the funds. It is also more neutral with regard to the number of transactions per dollar of deposits, which may be very sensitive to the location of branch. For instance, branches will have lower costs per dollar of deposits if they have customers with fewer transactions per dollar in their accounts. This may erroneously be measured as high efficiency in the intermediation approach. This problem does not arise in the production approach, since number of transactions is directly measured as a service output.

However, as argued in Berger and Humphrey (1997), "*Neither of these two approaches is perfect because neither fully captures the dual roles of financial institutions as (i) providing transactions/document processing services and (ii) being financial* *intermediaries that transfer funds from the savers to investors.*" This failure is attributed to the dual role of deposits. Deposits have the input characteristics since they are raised by banks as the raw materials for loans. However, deposits also have the output characteristics because they are associated with a substantial amount of liquidity, safekeeping, and payments services provided to depositors.

Some studies (Berger *et al.*, 2009, Carvallo and Kasman, 2005, Cavallo and Rossi, 2002, Fu and Heffernan, 2007, Hao *et al.*, 2001, Hasan and Marton, 2003, Kasman and Yildirim, 2006, Kraft and Tirtiroglu, 1998, Maudos *et al.*, 2002, Maudos and Pastor, 2001, Pastor and Serrano, 2005 and Lozano-Vivas, 1997) use the **dual approach** to capture the dual role of deposits. Interests paid on deposits are counted as part of costs and the rate paid is included as an input price, both consistent with the input of the raw material of investable funds. Values of deposits are specified as outputs.

A few studies also used the **value added approach** (Bonin *et al.*, 2005a, 2005b, Dietsch and Lozano-Vivas, 2000, Fries and Taci, 2005, Maggi and Rossi, 2003, Sensarma, 2006), under which, the choice of bank products to be included is based on the concept of value added. Those banking functions requiring significant expenditures on nonmonetary inputs such as labour and physical capital to produce noninterest banking services are identified as outputs. Therefore, commercial and industrial loans, consumer and real estate loans, total securities and core deposits are normally treated as output. The main argument here is whether core deposits should be treated as output. Berger and Humphrey (1992) estimate that the implicit revenues of all US banks account for approximately 82 percent of total deposit revenue. Thus, if all deposits service can be explicitly priced, core deposits would produce substantial service output.

2.1.2.2.2 Control for loan quality and risk factor

In theory, comparison of bank's performance should be conducted among banks with the same output quality. However, it is possible to have unmeasured differences in quality because the banking data may not fully capture the heterogeneity in bank output. Different types of loans may have different risk characteristics, even for the same type of loans. For example, commercial loans can vary in size, repayment schedule, risk, transparency of information, type of collateral, etc. These differences may not be captured by the output quantity in the model. Therefore, as suggested in Mester (1996) that "*Unless quality and risk are controlled for, one might easily miscalculate a bank's level of inefficiency*; *e.g., banks scrimping on credit evaluations or producing excessively risky loans might be labelled as efficient when compared to banks spending resources to ensure their loans are of higher quality.*" Hughes and Mester (1993) argued that the quality of bank's asset and the probability of banks' failure could influence a bank's cost in a variety of ways. First, if a bank holds a large proportion of nonperforming loans, this may signal that the bank uses fewer than the usual amount of resources in the initial credit analysis and followed monitoring procedure. Therefore, although lower quality loans may provide short-run cost savings, it may also entail extra administrative expenses for the bank as it tries to resolve these bad loans in the long run. Additionally, since quality of the bank's asset may influence the probability of its failure, the cost of debt (deposit) may also be affected. As suggested by Hannan and Hanweck (1988), the interest expenses of uninsured deposits contain a risk premium; the low asset quality can increase the interest cost of uninsured deposits for banks. For these reasons, Hughes and Mester (1993) and Mester (1996) include the average volume of nonperforming loans as a measure of loan quality. In Altunbaş *et al.* (2000), Huang (2000) and Mertens and Urga (2001), the non-performing loan ratio is used as a control for asset quality. Moreover, Hasan and Marton (2003) used loan loss provision to control for risk preferences, loan quality and ability to absorb losses.

However, Berger and DeYoung (1997) suggest that whether it is appropriate to include non-performing loans and loan losses in estimating the bank's cost function depends on the extent to which these variables are exogenous. Non-performing loans and loan losses would be exogenous if caused by negative economic shocks or unpredicted events ("bad lucks"), but they could also be endogenous, either because of the poor management in managing and monitoring the loan portfolio and controlling for the operating expenses ("bad management") or because of a conscious decision to reduce short-run expense by cutting back on loan origination and monitoring resources ("skimping") or for the incentives by increasing the riskiness of its loan portfolio, which results in higher non-performing loans on average in future ("moral hazard", a classical problem of excessive risk-taking when another party is bearing part of the risk and cannot easily charge for or prevent that risk-taking). As argued by Berger and DeYoung, *"Under the bad luck hypothesis, loan quality is driven by external events, and as such efficiency measurement should control for nonperforming loans in cost and profit functions. This would help remove by statistical means the extra costs of dealing with nonperforming loans—which were caused by bad luck, not by managerial inefficiency – rather than erroneously counting these extra costs as inefficiency. Under the bad management and skimping hypothesis, however, loan quality is driven by internal events. As a result, controlling for nonperforming loans in cost and profit functions will artificially increase measured efficiency by removing statistically the part of the cost inefficiencies (or revenue deficiencies) that are correlated with inefficient portfolio management. Neither hypothesis clearly dominates the other. Ultimately, whether or not one controls for loan quality should rest on the particular efficiency application at hand."* (Berger and DeYoung, 1997:15). Berger and Mester (1997) suggest use the ratio of non-performing loans to total loans in the bank's state since it is almost entirely exogenous and controls for bad luck in the bank's environment. While Fu and Heffernan (2007) exclude non-performing loans in the cost function for the reason that non-performing loans in Chinese banking system may be caused endogenously due to a poor risk management, cutting back on screening and monitoring, or making loan decisions without anticipating changes in the business cycle.

Another aspect of efficiency measurement is the treatment of financial capital. There are two reasons why financial capital should also be taken into account. First of all, it may influence the probability of banks' failure and interest costs. As known to all, the insolvency risk can affect a bank's cost and profit via the risk premium that bank needs to pay for the uninsured debt. At the meanwhile, the insolvency risk depends on the financial capital which can absorb the losses of nonperforming loans. Aside from concerns of risk, a bank's capital level will directly affect costs by providing an alternative funding source of loans as a substitute for deposits or other funding sources. Interests paid on debt counts are counted as cost, but dividends paid are not. On the other hand, raising equity capital involves higher costs than raising deposits. If the first effect dominates, measured costs will be higher for banks using a higher proportion of debt financing; if the second effect dominates, measured costs will be lower for those banks. Large banks depend more on debt financing to finance their portfolios than small banks do, so a failure to control for equity could yield a scale bias.

It is widely accepted (Altunbaş *et al.*, 2000, Altunbaş *et al.*, 2001, Hughes and Mester, 1993, Mester, 1996, Hao *et al.*, 2001, Hasan and Marton, 2003, Mertens and Urga, 2001, Vander Vennet, 2002 and Weill, 2004, 2003) that the level rather than the price of equity capital should be included since cost-minimization may not fully explain the level of equity capital. If a bank holds its equity level according to the objective of cost-minimization, price of equity capital can be used since the more expensive the price of equity capital, the fewer banks will choose to hold. However, in the real banking world, cost-minimization can only explain part of a bank's capital level. There are other factors bank manager should take into account when making the decision of the level of equity, for instance, the regulatory requirement and manager's preference to risk. Firstly, according to the Basel Accord, the minimum and the core capital ratio should be no less than 8% and 4%, respectively. These ratios are set with regard to the soundness of the whole banking system instead of individual bank's objective of cost-minimization. Therefore, it is possible that the regulations defining capital adequacy may constrain a bank to employ more financial capital than it would in an unregulated environment. This may eliminate the advantage afforded to banks of using deposit and debt financing since raising equity capital is much more expensive than raising deposit. Secondly, even if the regulations defining capital are not binding, a bank's level of financial capital may not be chosen to minimize cost if that level implies a degree of risk that is unacceptable. Hughes and Mester (1994) find evidence that banks exhibit nonneutrality toward risk and do not choose the cost-minimizing level of financial capital. A risk-averse bank might choose to fund its loans with a higher ratio of financial capital to deposits than a risk-neutral bank. Since the financial capital is typically more expensive than deposits, this might lead one to conclude that the risk-averse bank was producing its output in an allocatively inefficient manner (i.e., using the wrong input mix) where actually it is the risk-preferences that differ. Hence, allowing for the possibility of non-risk-neutrality suggests that the level rather than the price of financial capital should be included in the cost function. This also has the effect of ensuring that the estimated cost function refers to the short run rather than long run equilibrium.

The impact of quality factor and equity capital on the scale and scope economies and X-efficiency levels has been considered in the recent literature. Hughes *et al.* (1995) and McAllister and McManus (1993) find in US banking that either optimal bank size increases or scale economies are never exhausted when risk and quality variable included. As shown in Clark (1996), including equity capital in the measurement eliminates the scale diseconomies in production costs. Altunbaş *et al.* (2000) show that scale economies tend to be overstated if risk and quality factors are not taken into account since optimal bank size is considerably smaller when risk and quality factor are included in the cost function. However, empirical findings conflict in the impact of equity capital on cost efficiency, Clark (1996) finds that risk variables significantly alter X-inefficiency estimates for US banks, while Berger and Mester (1997) compare the variation with or without concerns about the quality factor and equity capital in the cost function and find little effect on the average level or dispersion of cost efficiency, although firms are ranked slightly differently. Altunbaş *et al.* (2000) also suggest that cost X-inefficiency appears less sensitive to risk and quality factors as well.

2.1.2.2.3 Environmental variables

One of the main assumptions in the frontier efficiency analysis is that all the producers share the same production technology and face the same environmental conditions. However, it may not be the case, especially in cross-country comparisons in which countries differ from each other in many aspects. It is widely suggested in the literature of international comparisons of banking performance, environmental variables should be included to control for cross-country heterogeneities. By applying the cross-sectional stochastic frontier framework, studies such as Carvallo and Kasman (2005), Dietsch and Lozano-Vivas (2000), Fries and Taci (2005), Kasman and Yildirim (2006), Pastor and Serrano (2005) and Perera *et al.* (2007) incorporate several vectors of environmental variables that reflecting the differences between countries' geographical, economic, and financial regulatory characteristics and suggest that control for cross-country differences is important as it may explain part of the inefficiency estimated when country differences are not controlled. However, this evidence is not appraised in panel data stochastic frontier framework that overcomes several shortcomings of cross-sectional framework. Detailed methodology review and comparisons of cross-sectional and panel data stochastic frontier framework will be discussed in next section 2.2.

2.1.2.3 The choice of functional form in efficiency measurement

Using parametric frontier techniques, either technical efficiency can be measured by using stochastic production frontier, or cost efficiency can be measured adopting stochastic cost frontier, or profit efficiency can be estimated employing stochastic profit frontier, or technical efficiency, for multiple output firm, can be obtained by utilizing stochastic distance frontier. However, no matter what kind of efficiency to be measured, a certain functional form has to be specified. The existing efficiency literature has witnessed the utilization of more flexible functional forms developed from the previously prevailing Cobb-Douglas functional form. Cobb-Douglas function is considered to be first-order flexible. However, *ceteris paribus*, one always prefers functional forms (such as translog function) that are second-order flexible, although it comes at the expense of more parameters to be estimated. As seen in my survey, only one study (Christopoulos and Tsionas, 2001) uses Cobb-Douglas function for simplicity to demonstrate their stochastic cost frontier with control for heteroscedasticity.

2.1.2.3.1 Translog functional form

Translog functional form is one of the most widely used functional forms in the empirical literature on bank efficiency. It presents the well-known advantages of being a flexible form. As argued by Coelli *et al.* (2005) and Kaparakis *et al.* (1994), it imposes few restrictions on the first- and second-order effects and at the same time, it can also be viewed as a second-order logarithmic approximation to an arbitrary continuous transformation surfaces. Therefore, translog function imposes few restrictions (by duality theory) on the production technology and in particular, it also envelops Cobb-Douglas specification. The general model of translog cost function is given by:

$$\begin{aligned}\ln C_{it} &= \alpha + \sum_{j=1}^{J}\beta_j \cdot \ln y_{jit} + \sum_{m=1}^{M}\delta_m \cdot \ln w_{mit} \\ &+ \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J}\beta_{jk} \cdot \ln y_{jit}\ln y_{kit} + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}\delta_{mn} \cdot \ln w_{mit}\ln w_{nit} \\ &+ \sum_{j=1}^{J}\sum_{m=1}^{M}\gamma_{jm} \cdot \ln y_{jit}\ln w_{mit} + u_{it} + v_{it}\end{aligned}$$

[2.1]

where a bank is supposed to use inputs $x_m \in R_+^m$ $(\mathrm{m}=1,...,\mathrm{M})$ available at fixed prices $w_m \in R_+^m$, to produce outputs $y_j \in R_+^J$ $(\mathrm{j}=1,...,\mathrm{J})$. Equity capital and non-performing loan ratio could be included in this model if the risk and quality factor are taken into account (see Hughes and Mester, 1993 and Mester, 1996). Since the duality theorem requires that the cost function is linearly homogenous in input prices and parameters of second-order partial derivatives are symmetric, the following restrictions have to be imposed to the parameters of the translog function form:

$$\beta_{jk} = \beta_{kj}\ \forall \mathrm{j,k},\ \delta_{mn} = \delta_{nm}\ \forall \mathrm{m,n},\ \sum_{m=1}^{M}\delta_m = 1, \sum_{m=1}^{M}\delta_{mn} = 0\ \forall\ \mathrm{m, and} \sum_{j=1}^{J}\gamma_{jm} = 0\ \forall\ \mathrm{j}$$

37 out of 48 studies in my survey adopt this flexible functional form to measure cost efficiency and scale and scope economies in US and European banking markets. The mean cost efficiency scores are relatively dispersed even if involved in the same banking market (see Table 2.1 PANEL A). It is not that surprising due to the distinctions among different objectives, different banking structures and regulatory forms, different input and output specifications and different sample sizes, etc. However, it is still worth summarizing their findings since inevitably these literatures may reach some convergence and concern about issues that may also be discussed in my own research.

First issue lies in whether large banks enjoy cost advantages over small banks. Theoretically, cost savings can be achieved through the expansion of output for the following reasons. Firstly, increased size could allow a more efficient organization of resources. For instance, specialization cannot be done in small banks due to limited sources. The same machines and workers must often be employed for a variety of tasks, and tellers may also be assigned to sort cheques and audit accounts. Large banks, however, may divide tasks so that employees and machines can be used in one facet of their operation. Thus, the productivity between both capital and labour arises with the scale of operations. Specialization could also result in a more economical use of materials purchased by the bank. Secondly, some types of technological innovations, such as computers, may be economically more feasible for large banks. Thirdly, due to the law of large numbers, large banks do not need to hold cash balances in the same proportion as smaller banks. Since holding cash balances is costly, larger banks should have lower costs for holding cash balances than their action demands. Moreover, larger banks are seemingly better able to diversify their assets and reduce risk as well as to offer various services to customers.

There is ambiguous evidence across the banking efficiency literature concerning the relationship between efficiency and size. Some empirical studies found the evidence supporting the argument that large banks are less efficient than small banks. Kaparakis *et al.* (1994) find the average cost efficiency reduced from 0.91 to 0.83 as bank size grows. Berger and Mester (1997) suggest that medium sized banks presented a slightly higher efficiency of 0.87 compared to 0.86 for large sized banks. Maggi and Rossi (2003) also find the slight 2% advantage from small banks to large banks. Allen and Rai (1996) find cost efficiency scores of 0.6 for large banks and 0.82 for small banks in both the US and European countries. Maudos *et al.* (2002) and Christopoulos *et al.* (2002) also find greater efficiency associated with small banks in Europe. Similar results have been found in China and transition countries. Berger *et al.* (2009) and Fu and Heffernan (2007) both find that the top four Chinese commercial banks are far less efficient than the joint-stock and foreign commercial banks. Bonin *et al.* (2005a, 2005b), Fries and Taci (2005), Kasman and Yildirim (2006), Kraft and Tirtiroglu (1998), Mertens and Urga (2001) and Weill (2003) all agree that privatized and foreign banks are more efficient than those large domestic state-owned banks. However, some studies do not find clear signs supporting the cost advantage for either small or big banks. Ferrier and Lovell (1990) use both parametric and non-parametric frontier techniques to measure the cost efficiency of US banks and find no signs of cost advantage enjoyed by large banks over small banks. Casu and Girardone (2002), Vander Vennet (2002) and Weill (2004) study the cost efficiency in European countries and they find that although banks' efficiency differs between size categories, no trend is apparent. Lang and Welzel (1996) use the fixed-effects model and random-effects model to measure the cost efficiency in German banking while the results from these two methods differ. The overall efficiency of these two models deviates dramatically from the optimal frontier. Under the random-effects model relative efficiency is lower but does not exhibit a size trend which can be observed under the fixed-effects model where small banks are more efficient than large banks. By controlling for scale differences, Kwan and Eisenbeis (1996) find that the average small banks are less efficient than the average large banks with a reported average efficiency of 0.81 for the former and 0.92 for the latter. Pastor and Serrano (2005) and Perera *et al.* (2007) also find that very large banks are slightly more efficient than small banks for both cost and profit efficiency.

The second issue concerns about the financial conglomerates, universal banks and foreign ownership. Casu and Girardone (2002) suggest that bank groups are more X-inefficient than bank parent companies and subsidiaries forming part of the groups. Vander Vennet (2002) suggest that under the criterion of conglomeration it can be observed that specialized banks are more efficient in traditional intermediation activities, while conglomerates appear to be slightly better managed when nontraditional activities are included. Also, under the criterion of universality, as reported by Molyneux *et al.* (1996), the sample of universal banks exhibits significantly higher average operational efficiency levels.

Some studies consider the impact of foreign ownership on the efficiency gains. Chang *et al.* (1998) investigate whether foreign-owned banks in the US operate more efficiently than the US-owned counterparty and the empirical results go against their expectations. It is widely accepted in efficiency literature of transition countries that involvement of foreign ownership usually brings greater efficiency improvement. This evidence is found in almost all transition efficiency studies (see Bonin *et al.*, 2005a, 2005b, Fries and Taci, 2005, Hasan and Marton, 2003, Kasman and Yildirim, 2006, Kraft and Tirtiroglu, 1998, Mertens and Urga, 2001 and Weill, 2003). Both Berger *et al.* (2009) and Fu and Heffernan (2007) support the argument of the impact of minority foreign ownership on improving banks' performance. Perera *et al.* (2007) find that banks with wide spread ownership are more cost efficient, while in Sensarma (2006) foreign banks are the worst performer compared to state-owned and private banks.

Furthermore, the efficiency analysis in the emerging markets is also worth studying because the characteristics of these countries, for instance, undergoing the transition of their economies, enforcing the deregulation of their banking system, and recovering from the economic crisis and rebuilding their banking system, may shed some light on my research aims. Carvallo and Kasman (2005) use the common cost frontier with country-specific environmental variables to estimate the cost efficiency of 481 banks from Latin American countries and find that the average level of inefficiency is 0.17 with the range from 0.09 for Honduras to 0.2 for Venezuela. The largest economies are found to be the most efficient. Hao *et al.* (2001) intend to identify the key determinants of Korean bank efficiency following the deregulation in the Korean banking sector and find that banks with higher rates of assets growth, fewer employees per million won of assets, larger amounts of core deposits, lower expense ratios and nationwide branches are more efficient. The average cost inefficiency of Hong Kong banks is found to be about 0.16 to 0.3 and to edge up following the financial crisis, which perhaps due to the additional resources spending to monitor non-performing loans (Kwan, 2002). Fu and Heffernan (2007) measure the impact of different ownership and banking reforms on Chinese banks' X-efficiency and their results show that on average, Chinese banks are operating 50% to 60% below the X-efficiency frontier. And as they expect, the joint-stock banks are, on average, about 8.5% more X-efficient than the state-owned banks. They argue that although major shareholder of joint-stock banks is the state, joint-stock banks, different from the state-owned banks that continue to assist in fulfilling social welfare objectives, are established to facilitate the development of an efficient banking system, and it is rare for them to be involved in the implementation of state policy. However, on the contrary, Berger *et al.* (2009) find that the state-owned banks are the most efficient as the big four state-owned banks are operating at the cost efficiency level of 0.84. It partly attributes this result to accounting practices or subsidies on the cost side for state-owned institutions.

2.1.2.3.2 Fourier flexible function

The translog function is developed as a local approximation to some unknown true underlying cost function. In practice, the translog function provides poor approximation to the true cost function when the global behavior of cost function differs from the local behavior. Thus, in its application, the translog function is potentially subject to misspecification. Compared with the translog function, the Fourier flexible cost function globally approximates the underlying cost function across a broad range of outputs. It has been widely accepted that the global property is important in banking where scale, product mix and other inefficiencies are often heterogeneous. Therefore, local approximations (such as those generated by the translog function) may be relatively poor approximation to the underlying true cost function.

The Fourier flexible function is a semi-non-parametric approach used to tackle the problem arising when the true functional form of the relationship is unknown. This methodology was first proposed by Gallant (1981, 1982) and applied to the analysis of bank cost efficiency by Berger *et al.* (1997) and Mitchell and Onvural (1996). It has desirable mathematical and statistical properties that a linear combination of the sine and cosine function, namely the Fourier series, can fit exactly any well behaved multivariate function. This is because the mathematical behaviour of the sine and cosine functions are mutually orthogonal over the [0, 2π] interval, so that each additional term can make the approximating function closer to the true path of the data wherever it is most needed.. In contrast, when using parametric methods like the translog, one holds the maintained hypothesis that the bank industry's true cost function is translog form. If this maintained hypothesis is false misspecification error occurs (see Mitchell and Onvural (1996) for brief discussion about the methodology).

The Fourier flexible function used in the banking efficiency studies usually includes a standard translog form and all first-, second-, third-order trigonometric terms, as well as a composed error structure, written as

$$
\begin{aligned}
\ln C_{it} &= \alpha + \sum_{j=1}^{J} \beta_j \cdot \ln y_{jit} + \sum_{m=1}^{M} \delta_m \cdot \ln w_{mit} \\
&+ \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J} \beta_{jk} \cdot \ln y_{jit} \ln y_{kit} + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} \delta_{mn} \cdot \ln w_{mit} \ln w_{nit} + \sum_{j=1}^{J}\sum_{m=1}^{M} \gamma_{jm} \cdot \ln y_{jit} \ln w_{mit} \\
&+ \sum_{p=1}^{P} \left[a_p \cos(z_p) + b_p \sin(z_q)\right] + \sum_{p=1}^{P}\sum_{q=1}^{P} \left[a_{pq} \cos(z_p + z_q) + b_{pq} \sin(z_p + z_q)\right] \\
&+ \sum_{p=1}^{P}\sum_{q=1}^{P}\sum_{s=1}^{P} \left[a_{pqs} \cos(z_p + z_q + z_s) + b_{pqs} \sin(z_p + z_q + z_s)\right] + u_{it} + v_{it}
\end{aligned}
\qquad [2.2]
$$

The model specification is the same as in translog cost function except the variable z_i in the sine and cosine terms. When considering the impact of risk and quality factor on the cost efficiency score, the environmental variables equity capital and the non-performing loan ratio can also be included. Thus **z** may represent the adjusted values of the log output and input prices, **y** and **w,** and log value of equity and non-performing loan ratio that span the interval $[0, 2\pi]$. Under this treatment, P will be equal to the sum of number of output, input prices and environmental variables (see Berger and DeYoung, 1997). However, Berger *et al.* (1997) leave the input prices effects to be defined entirely by the translog terms. The primary aim is to maintain the limited number of Fourier terms for describing the scale and inefficiency measures associated with differences in bank size. Moreover, the usual input price homogeneity restrictions can be imposed on logarithmic price terms; whereas they cannot be easily imposed on the trigonometric terms (also see Altunbaş *et al.*, 2001).

Since the duality theorem requires that the cost function is linearly homogenous in input prices and second-order parameters are symmetric, the following restrictions must be applied to the parameters of the translog function form:

$$\beta_{jk} = \beta_{kj} \; \forall \mathrm{j,k}, \; \delta_{mn} = \delta_{nm} \; \forall \mathrm{m,n}, \; \sum_{m=1}^{M} \delta_m = 1, \sum_{m=1}^{M} \delta_{mn} = 0 \; \forall \; \mathrm{m, and} \sum_{j=1}^{J} \gamma_{jm} = 0 \; \forall \; \mathrm{j}$$

Empirical studies conflict on whether Fourier flexible function dominates the translog function with the global approximation to the “true” cost frontier. Mitchell and Onvural (1996) find robust evidence that Fourier flexible function fits the data better. They suggest that the previous efficiency studies using the translog function may be misleading and bank policymakers who have been guided by these researches should rethink their decision. Berger *et al.* (1997) test of the null hypothesis that the nested translog specification is correct is rejected at the 1% level in 5 of 6 cases and at the 5% level in the other cases, supporting the Fourier flexible function. In Berger and DeYoung (1997), the average bank is measured to be about 92 percent efficient over the entire sample period, a higher level of cost efficiency is found in most other studies. They suggest that most of this difference reflect their more general specification of the cost function (Fourier flexible rather than translog) and the distribution of the inefficiency term (truncated normal distribution rather than the half normal). Maggi and Rossi (2003) use Fourier flexible and translog function in evaluating the robustness of the empirical results through several specifications proposed. Although estimated parameters present consistent values in different cost functions, tests performed on the specifications are in favour of the Fourier flexible function.

However, Berger and Mester (1997) find only a small difference in average efficiencies when substituting the translog form for the Fourier flexible function since the average efficiencies are lower by about 1% with about the same degree of dispersion. Altunbaş *et al.* (2001) find that the average inefficiencies in European banking appear to range between 20% and 25% across the different size classes. This finding is consistent with Molynuex, *et al.* (1996) and Vander Vennet (2002), which also indicate the little difference in using both the translog and Fourier flexible function. Carvallo and Kasman (2005) intend to use the Fourier flexible function but they fail to reject the null hypothesis of all Fourier parameters jointly equal to zero. Fu and Heffernan (2007) argue that the Fourier flexible specification requires more degrees of freedom but with only a few banks and a short history, the Chinese banking data are limited by comparatively few observations. Therefore, they adopt the translog function rather than the Fourier flexible function.

Another important aspect that the researcher who favours Fourier flexible function should pay attention to is the fact that the popularity of using the Fourier flexible function in the recent banking efficiency literature comes from its capability of providing a better fit to data. However, as tested in Altunbaş and Chakravarty (2001), the goodness-of-fit criterion is an unreliable indicator of forecasting ability, thus casting doubts on the validity of conclusions derived from the use of Fourier flexible function. Moreover, one way of addressing the issue of the local approximation property of the translog function is to carefully select the point of approximation. By using log mean corrected data in the sample the translog function is expanded around an approximation point at the sample mean therefore avoiding the danger of arbitrarily approximating at an outlier observation.

2.1.2.4 Economies of Scale and Economies of Scope

2.1.2.4.1 Economies of scale

The concept of scale economies refers to the rate at which output changes as all factor quantities are varied. Firms are said to show economies of scale if technology allows production cost to rise proportionately less than output when output increases. That is, economies of scale exist if long run per unit or average production costs decline as output rises. Conversely, if average costs rise with output, diseconomies of scale are present. Scale economies are based on the shape of the long run average cost curve, which illustrates average costs at each level of output. Figure 2.3 displays the long-run average cost (LAC) curve and the long-run marginal cost (LMC) curve with a series of short-run average cost (SAC) and short-run marginal cost (SMC) curves.

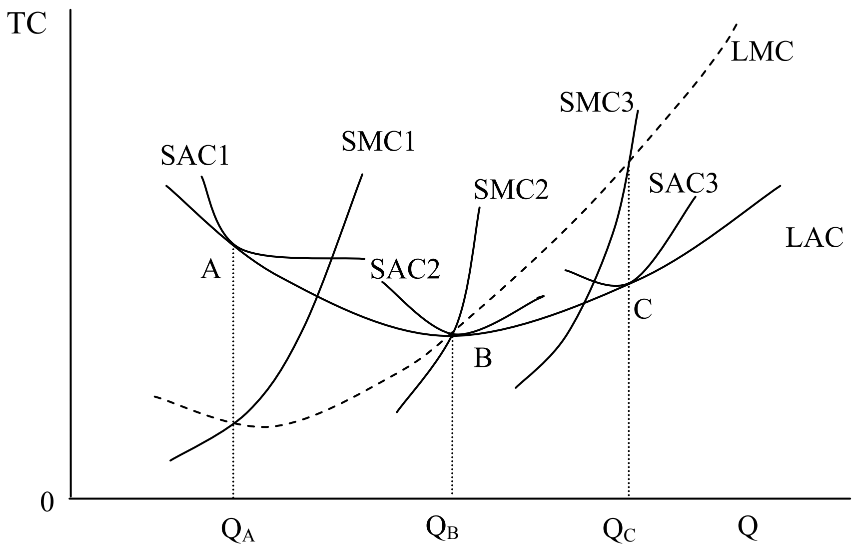

Figure 2.3: Economies of scale and the average and marginal cost curve shapes

The average cost curve shows the average cost per unit of output at different levels of output, and the marginal cost, the additional cost incurred when producing a very small increment of output is similar to the rate of change in average cost. Economies of scale are illustrated up to the output level Q_B, where the LMC curve lies below the LAC curve, and diseconomies of scale thereafter, where the LMC curve lies above the LAC curve. Scale economies in the single product firm can be measured as:

$$SE = \frac{MC}{AC} = \frac{\partial TC / \partial Q}{TC / Q} \qquad [2.3]$$

which is simply the elasticity of cost with respect to output. Therefore, SE<1, SE=1, SE>1 refers to the increasing returns to scale, constant returns to scale and decreasing returns to scale.

2.1.2.4.2 Economies of scope

Economies of scope arise if two or more products can be jointly produced at a lower cost than being produced separately in their independent production. Diseconomies of scope are presented if joint production is more costly than independent production. The concept of economies of scope can be explained geometrically in Figure 2.4.

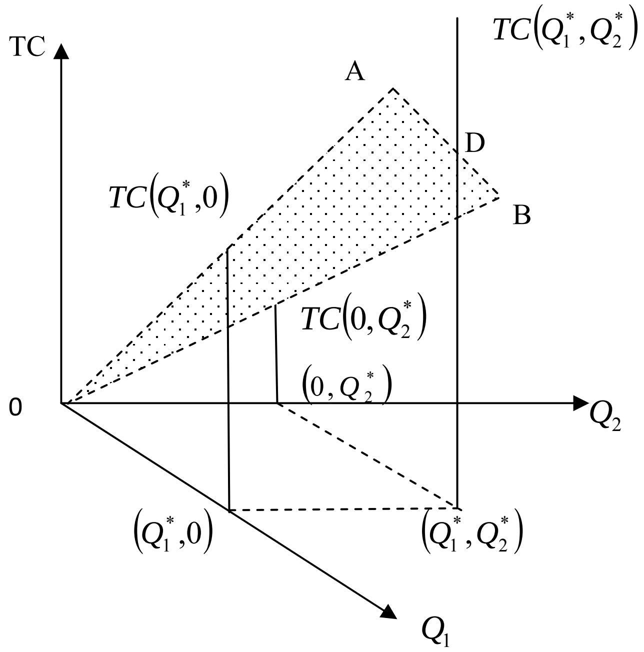

Figure 2.4: The concept of scope economies (adapted from Baumol *et al.*, 1988, p.72)

The figure illustrate that the concept of economies of scope involves a comparison of $TC(Q_1^*,0)+TC(0,Q_2^*)$, the sum of the heights of the cost surface over the corresponding points on the axes, with $TC(Q_1^*,Q_2^*)$, the height of the cost surface at point (Q_1^*,Q_2^*) which is the vector sum of $(Q_1^*,0)$ and $(0,Q_2^*)$. If $TC(Q_1^*,Q_2^*)$ lie below the hyperplane OAB which goes through the origin and point $TC(Q_1^*,0)$ and $TC(0,Q_2^*)$, then the condition of scope economies is achieved. The degree of economies scope can be measured as follows:

$$SCOPE = \frac{TC(Q_1)+TC(Q_2)+\ldots+TC(Q_N)-TC(Q_1,Q_2,..Q_N)}{TC(Q_1,Q_2,..Q_N)}$$

[2.4]

where TC(.) denotes the total cost of production. If SCOPE>0 then there are economies of scope. SCOPE<0 implies diseconomies of scope. Given multi-product industry structure of the banking sector, economies of scope, and the concept related to it, will play an important role in the efficiency analysis of the banking sector.

2.1.2.4.3 Sources of economies of scale and scope

The literature on the theory of banking firm has hypothesized numerous ways in which economies of scale and scope might arise in the production (Molynuex *et al.*, 1996 and Clark, 1988). Specialised labour, computer and telecommunications technology and information are regarded as the main sources of economies of scale and scope. Making better use of specialized labour and capital and spreading fixed costs over large levels of output are usually cited as the predominant sources of economies of scale. Most economies of scope are thought to arise from the joint usage of a fixed resource.

Firstly, for a small bank, labour input may be used for non-specialised activity meaning that it cannot be dedicated to highly skilled or specialised activities. Therefore, the unspecialized labour can be treated as a fixed input that can be shared in the production of a number of products, with the potential to create economies of scope. As these smaller institutions grow, they may be able to fully employ more specialized labour in producing some or all of their products. If the expertise of specialized labour results in the process of a greater volume of deposit and loan accounts per unit of labour, then per unit labour costs can be reduced through increased specialization. In this case, increase size will result in economies of scale.

Secondly, the adoption of computer and telecommunications equipment can provide another basis for both economies of scale and scope at banks. Despite the large set-up costs required, computer and other electronic funds transfer equipment can process a large volume of transactions at a small additional cost per transaction. As banks increase the number of transactions of all types that can be performed by this equipment, it may be possible to reduce the per-unit cost of the firm as a whole as well as for individual products. Embracing this technology may provide a basis of both overall and product-specific economies of scale. In addition, any excess capacity of the equipment could be used to process other types of accounts at a small additional cost per transaction, thus achieving economies of scope.

Furthermore, economies of scale and scope may also be generated by information production. Before lending decision can be made, credit information must be gathered and analysed. This information can be reused in other lending decisions to reduce the incremental costs of extending additional credit. If the information is reused to make similar loans to the same customer or to other customers in the same region or industry, it will provide a source of economies of scale. Alternatively, if the information can be used to make unrelated types of loans to the institution's customers, it may serve as a source of economies of scope.

2.1.2.4.4 Empirical evidence of scale and scope economies in banking markets

In the empirical studies of economies of scale and scope, measures of scale and scope differ due to different research objectives, different banking market involved, different model specification and different sampling data (see Clark, 1988, Appendix B for summary).

There are two excellent surveys of studies of economies of scale and scope in the US and European banking markets. Clark (1988) reviews 13 studies that attempt to estimate economies of scale and scope for credit unions, savings and loan associations and commercial banks. These studies suggest a few broad conclusions. First, economies of scale appear to exist at low level of output with diseconomies of scale at large output levels. Second, there is no consistent evidence of economies of scope associated with banks. Molyneux *et al.* (1996) summarize the early single-product cost studies and post-1965 multi-product cost studies in the US and the existing scale and scope economies studies in the European market. The general empirical results of studies using the Cobb-Douglas cost function indicate that economies of scale exist in small and medium-sized banks, whereas diseconomies of scale are present in larger banks. Empirical analysis using the translog functional form find the evidence of the existence of U-shaped cost curves despite the various efficient scale points. Studies of larger European banking markets suggest the substantial economies of scale for the small and medium-sized banks while evidence of economies of scope is uncertain, the same as in US banking markets.

Three US banking studies and four European studies (Berger *et al.*, 1987, Ferrier and Lovell, 1990, Gilligan *et al.*, 1984 for US, Casu and Girardone, 2002, Lang and Welzel, 1996, Maggi and Rossi, 2003, Vander Vennet, 2002 for EU) adopting the translog functional form report more or less similar results to those in Clark (1988) and Molyneux *et al.* (1996). Those findings suggest that bank cost is a U-shaped curve, indicating economies of scale in small banks while diseconomies of scale in large banks. If the true bank cost is U-shaped, one may ask why large banks still exist and merger activities between large banks occur all over the world. Or, whether the detection of U-shaped cost curve is due to misspecification in the cost function.

McAllister and McManus (1993) argue that previous studies in the economies of scale and scope may be misleading due to the misspecification of the statistical technology. They suggest that the choice of fitting the translog cost function over a population of banks that varies widely in terms of size and output mix is the main cause of the puzzling results. The globally fitted translog cost function systematically misrepresents cost for certain types of banks, resulting in a specification bias that contributes substantially to the conclusion of decreasing returns to scale among large banks. McAllister and McManus (1993) estimate separate cost function over the sub-dataset to test the capacity of translog cost function to represent the global behaviour of banks. Unfortunately, the specification bias is found, indicating that fitting a single translog cost function over the entire sample may result in an unreliable scale efficiency estimates for large banks. Moreover, they believe that the misspecification may also come from neglecting the impact of equity capital and loan quality factor in the model.

Hughes and Mester (1993) and Mester (1996) take the equity capital into account and their empirical results suggest constant returns to scale all across the sample size and no evidence of economies and diseconomies of scope. Empirical studies using the globally fitted functional form such as Fourier flexible cost function and kernel functional form find distinct results. McAllister and McManus (1993) adopt the non-parametric kernel regression technique and include equity capital in the model. They find substantial increasing returns to scale in banks with a total asset size below $500 million. Above that point, there appears to be approximately constant returns to scale. This finding is consistent with that of Altunbaş *et al.* (2000), which use Fourier flexible function with equity capital included. The empirical results show that scale economies are prevalent in small banks. While in the dominant economies in the European, such as Germany and UK, large banks are more likely to show constant returns to scale. Mitchell and Onvural (1996) compare the estimation of the translog function form with that of the Fourier flexible form using the bank data in 1986 and 1990. The results from the Fourier flexible form suggest that banking in the smallest size group enjoy increasing returns to scale while banks in other size groups face constant returns to scale. However, results from the translog function suggest that increasing returns to scale spread over a wider range for the first four size group while constant returns to scale and slight economies of scope are found in the largest size group.

However, two conflicting results are also presented in the literature where both the Fourier flexible function and translog function are adopted. Altunbaş *et al.* (2000) suggest that scale economies are prevalent for all but largest Japanese banks when taken the risk factor into account, only smallest banks exhibit significant scale economies with the majority of banks experiencing diseconomies of scale. This finding is consistent with the ones using translog cost function. This conflicting result could be related to a variety of factors including: different institutional structures, the decline in capital strength of Japanese banks during the 1990s, the different time frame covered and growth variation in the output/input mix. Berger and Mester (1997) find substantial economies of scale across all size categories. However, when they try to figure out whether this powerful finding is due to the adoption of Fourier flexible form and consideration of equity factor, an interesting result occurs. When estimating using the translog functional form, even better results are found. Also, although re-estimation without specifying equity capital does reduce the measured scale economies for large banks, it does not eliminate them. It must be some power which is far more beyond than that of methodology that creates the better results such as the environmental changes in the 1990s in US banking sector, such as the lower open-market interest rate, regulatory changes and improvements in technology, etc. Therefore, as mentioned before, there is no "best" methodology which can fit all the cases. Researchers have to pick what is thought to be the best methodology for the studies after considerable investigation of own sake.

Another finding in the literature is also worth investigating. Berger *et al.* (1987) find substantial diseconomies of scope for large banks. Under the intermediation approach, measured diseconomies of consolidation are about one to three percent, while these effects are slightly higher under the production approach, suggesting that dividing the bank's output among two smaller firms may generate about one to three percent cost savings. However, the product mix of these smaller firms may not be configured so as to meet the consumer's demand or efficiently reduce the risk through portfolio diversification. Therefore, this result is consistent with the hypothesis that banks in branching states choose output taking into account of customers' convenience, joint demand and risk consideration, which can increase the banks' costs but also increase bank revenue. However, these effects cannot be captured by the statistical cost functions. Thus, the total economies from joint production may be understated in the empirical estimates here and in other studies.

2.1.2.5 Efficiency estimates

As the main objective of efficiency studies is to inform the government or regulatory authorities of their policy effects and to assess bank managers' ability to achieve optimum in production, the judgement depends on the efficiency estimates. As the centre of efficiency studies had been scale efficiency or scale and scope economies in the 1980s, it has been switched to cost and profit efficiency nowadays.

Cost efficiency measures the extent to which a bank's performance is close to the best practice banks for producing the same outputs under the same environmental conditions. That is, it measures how close is the bank's cost to the minimum cost determined by the best practice banks in the sample. It is derived from estimating a cost function in which total costs is regressed as a function of outputs, price of inputs, environmental variables, random noise and inefficiency, written in log terms as

$$\ln C_{it} = \ln f(\mathbf{y}, \mathbf{w}, \mathbf{z}) + v_{it} + u_{it} \qquad [2.5]$$

where C_{it} measures total costs, **y** and **w** are the vector of outputs and input prices, respectively, **z** is the vector of environmental variables that may affect banks' performance, v_{it} denotes the random noise that may include measurement error, sampling error and luck that could temporarily cause higher or lower costs, and u_{it}

denotes cost inefficiency that drives banks away from the efficiency frontier and raise banks' costs over the best practice level. Cost efficiency is measured as ratio of actual costs to the minimum costs produced by the best practice bank under the same output level and conditions.

Two kinds of profit efficiency measures exist in the literature. One is the standard profit efficiency and the other is the alternative profit efficiency. Standard profit efficiency measures how close a bank is to achieve the maximum profit given a certain level of input prices and output prices. In contrast to cost efficiency measure the standard profit efficiency measure allows consideration of revenues that can be earned by varying outputs as well as inputs. The standard profit function, written in log terms, is

$$\ln(\pi_{it} + \theta) = \ln f(\mathbf{p}, \mathbf{w}, \mathbf{z}) + v_{it} + u_{it} \tag{2.6}$$

where π_{it} is the variable profits of banks measured as revenues of interest and fee incomes minus variable costs; θ is added as a constant to ensure the natural log is taken on a positive number; p and w are the vectors of output prices and input prices, respectively, while v_{it} and u_{it} are random noise and inefficiency. Standard profit efficiency is measured as ratio of actual profits to the maximum possible profits earned by the best practice bank in the sample. A standard efficiency ratio of 0.80 indicates that the bank is losing 20% of its profits that could be achieved because of excessive cost used or insufficient revenue raised.

As argued by Berger and Mester (1997), profit efficiency concept provides a better measure than cost efficiency concept when evaluating banks' overall performance. Cost efficiency accounts for errors only on input side while profit efficiency takes accounts of errors not only on input side but also on output side. Cost efficiency is based on economic objective of cost minimization that requires bank manager to focus on reducing operating costs. However profit efficiency is based on a more accepted economic objective of profit maximization under which bank managers need to pay an equal amount attention to raise marginal revenue as to reduce marginal cost. In this sense, a bank that is more profit efficient by earning positive marginal profit may be inappropriately measured as less cost efficient. Moreover, cost efficiency evaluates performance at a given level of output, which may not necessarily to be the optimal level, whereas profit efficiency examines banks' performance based on comparisons with the best practice of profit maximization, which corresponds to the optimum. Therefore, standard profit efficiency may take better account of cost efficiency against the optimal level than cost efficiency itself.

The alternative profit efficiency measures how close a bank is to achieve maximum profit at a given output level rather than output prices in standard profit efficiency concept. The alternative profit function uses the same dependent variable as in the standard profit function [2.6] but the same independent variables as in the cost function, written in log terms as

$$\ln(\pi_{it} + \theta) = \ln f(\mathbf{y}, \mathbf{w}, \mathbf{z}) + v_{it} + u_{it} \qquad [2.7]$$

which is identical to standard profit function except that y replaces p in the function and different $v_{a\pi}$ and $u_{a\pi}$ are estimated. The alternative profit efficiency is also measured as ratio of actual profit to the possible maximum profit earned by the best practice.

As suggested in Berger and Mester (1997: 901-904), alternative profit efficiency measure may be useful when one of the following conditions hold. First, if there are unmeasured differences in output quality of banking services, alternative profit efficiency provides a better measure than cost efficiency since alternative profit efficiency measure acknowledges the fact that additional revenue earned by providing high quality banking services may be good enough to offset the additional cost induced and high quality banks may be more profit efficient rather than mistakenly be judged as cost inefficient. Second, when banks cannot achieve every output scale, alternative profit efficiency may be more helpful since it can reduce the scale bias presented in the standard profit efficiency measure. Large banks and small banks differ in the size of output level. Since exogenous variables used in standard profit efficiency measure cannot tell the size difference, large banks may be termed as more profit efficient than small banks just because small banks cannot reach their output level. This scale bias can be controlled by alternative profit efficiency measure since it compares the banks' performance based on the same output level. Third, if output market is not completely competitive, banks may have the market power to set the prices that clear out its output. An optimizing bank can increase its revenue-cost margin by increasing service quality because there may not be enough competition. It can also choose to reduce its service quality to save more cost. The alternative profit efficiency provides a better measure to these optimizing choices and can be treated as a robustness test for standard profit efficiency that considers prices as fixed and allow output to vary freely. Finally, practically, if output prices are not measured correctly or information on output prices is missing, the alternative profit efficiency will certainly be more appropriate than standard profit efficiency measure.

Empirical evidence of cost and profit efficiency studies differs with regard to banking industries reviewed (see Table 2.1 PANEL A for cost efficiency studies and PANEL B for profit efficiency studies). In terms of cost efficiency, by employing different functional forms and estimation techniques, efficiency studies in US banking sector found the average cost efficiency scores in a range of 0.74 and 0.94 for a data period from 1986 to 1995 (Allen and Rai, 1996: 0.76; Berger and DeYoung, 1997: 0.92; Berger *et al.*, 1997: 0.94 for intermediation approach and 0.79 for production approach; Berger and Mester, 1997: 0.87; Chang *et al.*, 1998, 0.77; Ferrier and Lovell, 1990: 0.74; Kaparakis *et al.*, 1994: 0.9; Kwan and Eisenbeis, 1996: 0.8-0.9; Maudos and Pastor, 2001: 0.97; Mester, 1996: 0.91-0.92). While using data from 1995 to 1998, Maggi and Rossi (2003) reported a lower cost efficiency level of 0.62 and 0.63 when using Fourier flexible function and translog function respectively. Cost efficiency studies in major European countries provided a rather mixed result. In Allen and Rai (1996) studies, 12 European banking industries were examined and the overall efficiency was about 0.75 to 0.9 for a period from 1988 to 1992. With a larger data ranging from 1989 to 1997, Altunbaş *et al.* (2001) find the overall average cost efficiency around 0.75 to 0.8 across different asset size, which is similar to the efficiency of 0.79 from 1989 to 1996 in Carbo *et al.* (2004). Casu and Girardone (2004) look into five major European countries, and find the overall efficiency of 0.8 to 0.9 from 1993 to 1997. Their results are consistent with Maudos *et al.* (2002), in which 10 European banking industries have an efficiency of around 0.84 from 1993 to 1996, and with Pastor and Serrano (2005) where the authors report an efficiency level of 0.88 for 7 EU countries. Similar findings are also shown in Weill's (2004) study for 5 EU countries from 1992 to 1998 with average efficiency around 0.78. Moreover, Vander Vennet (2002) finds an overall efficiency level of 0.8 for 17 EU countries in year 1995 and 1996.

It is also interesting to see whether different studies covering the similar range of data period would provide the similar efficiency estimates for same country. First, I compare the results from Altunbaş *et al.* (2001) with those from Carbo *et al.* (2004). Both studies had similar efficiency estimates for all common countries except for Sweden (0.69 in Altunbaş *et al.*, 2001 but 0.83 in Carbo *et al.*, 2004). The second comparisons are among studies of Casu and Girardone (2004), Maudos *et al.* (2002), Pastor and Serrano (2005) and Weill (2004). Efficiency levels for Germany, Italy, Spain and UK were quite similar however, efficiency for France differs a lot (0.71 in Weill, 2004, 0.79 in Maudoes *et al.*, 2002, 0.83 in Casu and Girardone, 2004 and 0.93 in Pastor and Serrano, 2005). The higher efficiency score in Pastor and Serrano (2005), compared to other studies, may partly be explained by their sample selection since the number of French banks in their sample only accounts for half of those in other studies. Empirical findings of studies in transition countries are quite similar. The average efficiency are found around 0.78-0.8 when controlling for country differences and foreign banks are more efficient than domestic banks.

In terms of profit efficiency, Berger and Mester (1997) measure both the standard and alternative profit efficiency for 6000 US banks. The standard profit efficiency was around 0.55 and alternative profit efficiency was about 0.46. Maudos and Pastor (2001) also use alternative profit efficiency concept and found when using profit before tax as dependent variable, the estimated profit efficiency is around 0.55 for US banks and 0.45 for European banks. Similar alternative profit efficiency scores are also found in Maudos *et al.* (2002) and Pastor and Serrano (2005). Maudos *et al.* (2002) study 10 European banking industries from 1993 to 1996 and find the overall profit efficiency around 0.45. Pastor and Serrano (2005) study 7 European banking industries from 1993 to 1997 and report an average profit efficiency of 0.52. However, different profit efficiency scores are also found in European banking studies. Casu and Girardone (2002) also measure the alternative profit efficiency by using profit before tax as dependent variable but find the profit efficiency around 0.75. This finding is consistent with Vander Vannet (2002) in which the author finds profit efficiency of 0.69. The main reason behind these different results may be partly explained by the choice of input and output specifications. Maudos and Pastor (2001), Maudos *et al.* (2002) and Pastor and

Serrano (2005) all use dual approach that specifies deposits as both output and input, while Casu and Girardone (2002) and Vander Vennet (2002) adopt the intermediation approach that considers deposits as an input only. Profit efficiency studies in transition countries share the same conclusion of cost efficiency studies in those countries that foreign banks are more efficient than domestic banks.

2.1.2.6 Determinants of efficiency

Another important aspect in banking efficiency studies is to identify the sources of inefficiency. It has been suggested that acknowledging the potential sources causing inefficiency in banks' operation can help make appropriate policies with clear macroeconomic and regulatory implications. Majority of banking efficiency studies address this issue by applying a two-stage estimation approach. The first stage involves specification and estimation of a stochastic production function and prediction of cost (or profit) inefficiency (or efficiency) score for individual bank. Then, in the second stage, inefficiency (or efficiency) will be regressed on several explanatory factors. However, the two-stage formulation suffers two serious econometric problems. First, it must be assumed that the explanatory variables in the second stage are uncorrelated with independent variables such as output, input prices and environmental variables. If they are correlated, the maximum likelihood estimates of parameters in the first stage will be biased because of the omission of those explanatory variables from the stochastic frontier model. Consequently the estimated efficiencies being explained in the second stage are biased estimates from the true efficiencies, as they are estimated by the biased presentations of production frontier model. Therefore, even the "best" two-stage models will contribute nothing to the determinants of efficiency variation. Second, an inconsistency occurs in the above two stage method. As argued by Battese and Coelli (1995) and other researchers, the stochastic frontier production function is estimated under the assumption that inefficiency term is identically distributed, while in the second stage, the predicted inefficiency (or efficiency) term is regressed upon several environmental variables, suggesting that the inefficiency term is actually not identically distributed. In other words, for instance, in the first stage, the mean of inefficiency term is constant ($E(u_i) = \sqrt{2/\pi} \cdot \sigma_u$ in half normal case and $E(u_i) = \sigma_u$ in the exponential case), while in the second stage, the mean of efficiency term depends on some specification of explanatory variables. Therefore, a more appropriate approach involves a model specification in which both relations are estimated in a single stage. The most popularly used single-stage stochastic cost frontier is Battese and Coelli (1995), which can be written as

$$\begin{aligned} \ln C_{it} &= f\left(\mathbf{y}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}\right) + v_{it} + u_{it} \\ v_{it} &\sim N\left(0, \sigma_v^2\right) \\ u_{it} &= u(\mathbf{k}_{it}) \geq 0 \end{aligned}$$

[2.8]

where u_{it} is a function of a vector of exogenous explanatory variables **k** that are directly estimated simultaneously with other parameter estimates of output, input prices and possible environmental variables. An explicit specification of $u(\mathbf{k}_{it})$ used in Battese and Coelli (1995) is that u_{it} are independently (but not identical) distributed as non-negative truncations of a general normal distribution of the form:

$$N\left[\delta_0 + \sum_{k=1}^{K} \delta_k k_{kit}, \sigma_u^2\right]$$

[2.9]

where δ_0 and the δ_k are parameters to be estimated.

Within this framework, parameters $\left(\beta_0, \beta_k, \delta_0, \delta_k, \sigma_u^2, \sigma_v^2\right)$ are obtained by using MLE, which involves the reparameterisation $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\gamma = \sigma_u^2 / \sigma^2$. The next step is to obtain the technical efficiency estimates by using the Battese and Coelli (1988) expression of conditional expectation of $\exp(-u_{it})$, given ε_{it}. This is equal to

$$TE_i = E\{\exp(-u_{it}) \mid \varepsilon_{it}\} = \left[\frac{\Phi(\mu_{it} / \sigma_* - \sigma_*)}{\Phi(\mu_{it} / \sigma_*)}\right] \cdot \exp\left\{-\mu_{it} + \frac{1}{2}\sigma_*^2\right\}$$

[2.10]

where $\Phi(.)$ denotes the distribution function of the standard normal random variable,

$$\mu_{it} = (1-\gamma)\left[\delta_0 + \sum_{j=1}^{M} \delta_j k_{j,it}\right] - \gamma \varepsilon_{it} \quad [2.11]$$

and $\sigma_* = \gamma(1-\gamma)\sigma^2 = \sigma_u^2 \sigma_v^2 / \sigma^2$

Table 2.2 summarizes the explanatory variables used to examine the determinants of cost (or profit) inefficiency (or efficiency). Explanatory variables usually cover the categories of bank own characteristics, geographical, organizational and regulatory characteristics. Although there is no consensus of what variables should be used in analysis, some common factors are usually included such as bank size, ROA, ROE, equity ratio, non-performing loan ratio and ownership dummies. Therefore, it is worth summarizing the findings of these factors. First is the bank size effect. Allen and Rai (1996) find the significant positive size factor indicating the larger inefficiency associated with large banks. Similar results are found for the Chinese banking sector (Berger *et al.*, 2009), in European banks (Girardone *et al.*, 2004, Maudos *et al.*, 2002) and in the US banks (Kaparakis *et al.*, 1994). Mertens and Urga (2001) and Mester (1996) both find the coefficient of size effect is insignificantly different from zero. However, Altunbaş *et al.* (2000) find significant negative size effects suggesting that inefficiency decreases with bank growth. Carvallo and Kasman (2005) exploited the similar significant negative size effect in Latin American banking industries. Same evidence has been presented in Cavallo and Rossi (2002) for banks in Germany, Netherlands, Spain and UK, in Christopulos *et al.* (2002) for Greek banks, in Hasan and Marton (2003) for Hungarian banks and in Perera *et al.* (2007) for four South Asian banking sectors.

Second, for effects of ROA and ROE, it is expected that banks with great efficiency will have higher ROA and ROE. Not surprisingly, all the empirical studies considering ROA as a factor to explain the cost efficiency (or inefficiency) reach the consensus of favouring the initial expectation, despite different banking data set in different countries are examined using different efficiency techniques and models. For example, Allen and Rai (1996) find that the coefficient of ROA is significantly negative for large banks in both universal and separated banks, indicating that for larger banks, great cost efficiency is connected with higher profitability. Similar findings are provided by Altunbaş *et al.* (2000) for Japanese banks, Carvallo and Kasman (2005) in Latin American banking sector, Cavallo and Rossi (2002) and Weill (2004) in EU countries, Christopoulos *et al.* (2002) for Greek banks, Mertens and Urga (2001) for Ukraine banks, and Perera *et al.* (2007) for South Asian banks. The supporting argument is strengthened by Mester (1996), in which the author provides evidence not only on cost efficiency but also on profit efficiency of which the coefficient is significantly negative suggesting banks with low profit inefficiency present higher profitability.

Third, equity capital ratio and non-performing loan ratio are examined as controls for output quality and risk factor. The impact of considering these two factors in efficiency estimation has been discussed in section 2.2.2.2.2. The literature that incorporates the equity capital ratio reaches no accord about its impact on banks' performance. Allen and Rai (1996) find that increase in equity capital ratio will have a significant negative effect on cost efficiency for universal banks and large banks in separated banking countries only. While Altunbaş *et al.* (2000) indicate a strong positive effect of equity capital on cost efficiency instead. Their argument is supported by Carvallo and Kasman (2005) in Latin American countries and Hasan and Marton (2003) in Hungry. Cavallo and Rossi (2002) provide mixed evidences in EU countries as they find significant a positive relationship between equity capital ratio and cost inefficiency for Germany and Italy but a significant negative relationship for UK. Girardone *et al.* (2004) also find significant negative effects of equity capital ratio on banks' inefficiency suggesting that higher equity capital ratio would reduce banks' insolvency risk and controlling banks' operating risk that would bring better performance in banking service. Similar negative relationship is also found in Kaparakis *et al.* (1994) and Perera *et al.* (2007). Hao *et al.* (2001) and Mertens and Urga (2001) found the effect of equity capital on banks' inefficiency is insignificantly different from zero. Unlike equity capital ratio, no shocks are provided in the evidence of positive effects of non-performing loan on inefficiency as all four studies reported significant relationship between non-performing loan and inefficiency (see Altunbaş *et al.*, 2000, Carvallo and Kasman, 2005, Girardone *et al.*, 2004, Mertens and Urga, 2001).

The fourth effect, ownership effect on efficiency examined in emerging and transition countries is usually expected to have the positive effect of privatization and involvement of foreign ownership. Bonin *et al.* (2005a, 2005b) look into the ownership effect on banking performance in transition countries. They find the expected significant positive effect of foreign ownership on both cost and profit efficiency and insignificant negative effect of state-ownership. Weill (2003) also finds positive foreign ownership impact on cost efficiency and the author suggests that the gap between domestic-owned banks and foreign-owned banks is independent of bank size.

Consistent with the above results, Hasan and Marton (2003) find the foreign involvement in Hungarian banks is significantly associated with low inefficiency. By using the disaggregated foreign ownership variables, the authors find strong inverse relationships between inefficiency and foreign ownership share but insignificant effect of the minority foreign ownership. Berger *et al.* (2009) study the cost efficiency in Chinese banking system and they find that state-owned banks are the least efficient and foreign banks are the most efficient and suggest that the minority of foreign ownership is associated with greatly improved efficiency. The same evidence is provided by Fu and Heffernan (2007) in which the authors find positive significant coefficient of ownership dummy (one for state-owned, otherwise zero) suggesting joint-stock banks are more efficient than state-owned bank. Perera *et al.* (2007) used dummy variable that equals to one if the bank is state-owned and reported a statistically significant coefficient of 0.59, indicating that state-owned South Asian banks are more cost inefficient than private-owned ones. Similar results are also found in Italian banking sector (Girardone *et al.*, 2004).

Table 2.2: Explanatory variables used as determinants of inefficiency (or efficiency)

	Bank characteristics	**Geographical and organizational characteristics**	**Other characteristics**
Allen and Rai (1996)	ROA, TC/TA, SE/TA, Deposits/TA,LOANS/TA, Bank size		
Altunbaş (2000)	Bank size, E/TA, ROA, Loans/TA, Off-balance sheet items/TA, Customer and short-term funds/total funds, Liquid assets/TA, NPL/total loans	Institutional dummy	
Berger *et al.* (2009)	Ownership dummies, Bank size dummies		
Berger and Mester (1997)	Bank size dummies, Age, Loan/TA, Dummies for nominal value of bank's swaps, futures and similar contracts exceeds 5% of total assets, Purchased funds/TA, Standard deviation of banks' ROA, Standard deviation of banks' ROE	8 dummies for organizational form,	Concentration, Share of local market deposits, Location dummy, Real state income growth, State unemployment rate, Other 9 dummies
Bonin *et al.* (2005a)	Ownership dummies, Bank size dummies		
Bonin *et al.* (2005b)	Ownership dummies		
Carvallo and Kasman (2005)	ROA/TA, TC/TA, SE/TA, ROA, DEP/TA,LOANS/TA, Bank size, Non-interest income/total income, NPL/total loans, Loan loss provision/total loans		
Cavallo and Rossi (2002)	Bank size, Bank size dummies, Liquidity, Asset items, Liabilities, Asset item composition, ROA, Non-interest income/total asset, Interest income/total income, Cost/Income, E/TA	Institutional dummies	
Chrisopoulos *et al.* (2002)	Bank size, ROA, Loan/Outputs, Investments/Outputs		
Fu and Heffernan (2007)	Ownership dummies, Purchased funds/TA, Loans/TA, Investments/TA, Non-interest income/Pre-tax profits	Reform dummies	
Girardone *et al.* (2004)	Bank size, Interest margin/TA, Number of branches, Customers loans and deposits/TA, Ownership dummies, NPL/TA, E/TA, Income/E	Location dummies, Instituional dummies	

Table 2.2: Explanatory variables used as determinants of inefficiency (or efficiency) (continued)

	Bank characteristics	Geographical and organizational characteristics	Other characteristics
Hao *et al.* (2001)	Age, Growth rate, STA (Salaries to asset ratio), Square of STA, Branches to deposits ratio, total employees/TA, Demand deposits/total deposits, Non-interest income/operating profits, E, Reform dummies	Location dummies, Reform dummies	
Hasan and Marton (2003)	Liquidity, Short-term loan/TA, Investments/TA, customer loan/TA, E/TA, Cost inefficiency score, Bank size, Age, Hours of service, Ownership dummies, Merger dummies		
Kaparakis *et al.* (1994)	Bank size, E/TA, Purchased funds/total deposits, Nonaccrual loans/total loans, Capital-labour ratio, Number of foreign branches, Number of households, Number of saving institutions, Population density, Deposits ratio, Asset ratio, Per capita income, Per capita deposits	Location dummies,	
Mertens and Urga (2001)	Bank size, Bank capital, NPL/total loans, TC/TA, ROA, ROE		
Mester (1996)	Age, Number of Branches, Bank size, Qualifying capital, ROA, Uninsured deposits/total deposits, construction and land development loans/total loans, real estate loans/total loans, customers loan/total loans,	Location dummies, Institutional dummies	
Maudos *et al.* (2002)	Bank size, specialization, other characteristic to specific bank	Market characteristics	
Perera *et al.* (2007)	Crisis dummies, Bank size, Ownership dummy, Dummy for listing, E/TA, ROA, Non-earning assets/TA		
Weill (2003)	Ownership dummy, Loan/Investment assets, Deposits/TA		
Weill (2004)	TC/TA, TA/Income, ROA, ROE		

2.2 Methodology review of frontier efficiency measurement

2.2.1 Introduction

In this subsection, a methodological review of frontier approaches to efficiency measurement will be provided. Frontier efficiency analysis origins from Farrell (1957), in which the author introduced a measure of economic efficiency as the product of technical efficiency and allocative efficiency based on a benchmark isoquant. Inspired by Farrell's idea, scholars have developed non-parametric and parametric frontier approaches to either calculate or estimate firms' efficiency with numerous applications to different industries. The focus of this methodological review is on the development of stochastic frontier approach, the most popular and applied parametric frontier approach in the efficiency literature. This subsection is organized as follows. Section 2.3.2 introduces the origin of frontier efficiency analysis with detailed illustration of Farrell's (1957) input oriented efficiency measure. Section 2.3.3 describes briefly the non-parametric frontier approach data envelopment analysis and its shortcomings that lead to my main discussion of parametric frontier analysis discussed in section 2.3.4. Introduction of deterministic frontier approach and stochastic frontier approach is provided in section 2.3.4.1 and 2.3.4.2 respectively. Introduction of stochastic frontier analysis is divided into two parts. Early development of cross-sectional stochastic frontier models will be discussed in section 2.3.4.2.1 while recent expansion in panel data stochastic frontier framework is demonstrated in section 2.3.4.2.2.

2.2.2 Origin

The modern literature of efficiency analysis started from Farrell (1957). Farrell is highly influenced by Koopmans's "technical efficiency" measure (Koopmans, 1951) and Debreu's "coefficient of resource utilization" (Debreu, 1951). In Koopmans's analysis, a producer is said to be technically inefficient when it can produce the same level of output with less amount of input, or can use the same package of input to produce more amount of output. This can be seen as the twofold orientation of technical component of economic efficiency, which is output augmenting and input contracting. The measure of technical efficiency introduced in Debreu (1951), initially termed as the coefficient of resource utilization, is defined as one minus maximum equiproportionate reduction, i.e. radial reduction of all inputs that still allows the production process to be continued. In his seminar paper, Farrell introduced a method to decompose the overall efficiency of a production unit into its technical and allocative components. A firm is said to be inefficient by producing less than maximum output from a given set of inputs or using more than the minimum input required for a given level of output (technically inefficient) or by utilizing the wrong mix of input given their prices (allocatively inefficient).

The efficiency analysis carried out by Farrell (1957) can be explained in Figure 2.5. Under constant returns to scale, a firm is considered to produce a single output by using two inputs, which labelled as P in the diagram. The unit isoquant SS' in the space (X1/Y, X2/Y) is termed as the minimum combinations of two inputs needed to produce a unit of output. Under this framework, every combination along the isoquant (like point Q and Q') is considered as technically efficient while any point above or to the right of it defines a technical inefficient producer since it can contract the use of inputs without reducing the output level. Isocost line AA', the slope of which equals the ratio of two input prices, measures the minimum cost to secure unit output. Considering firm P, it utilizes the input vector as shown in the graph to produce the unit output. It suffers from two kinds of inefficiencies: first, it is inefficient in the technical sense, since by moving from P to Q, it can produce the same amount of output with less inputs. Therefore the technical efficiency of firm P is given by the ratio of OQ/OP. Second, it suffers allocative inefficiency, since by moving from Q to Q', one could produce the same output at a lower cost level by adjusting its input levels until the ratio of marginal products equals the ratio of input prices. Therefore, allocative efficiency is measured by the ratio of OR/OQ.

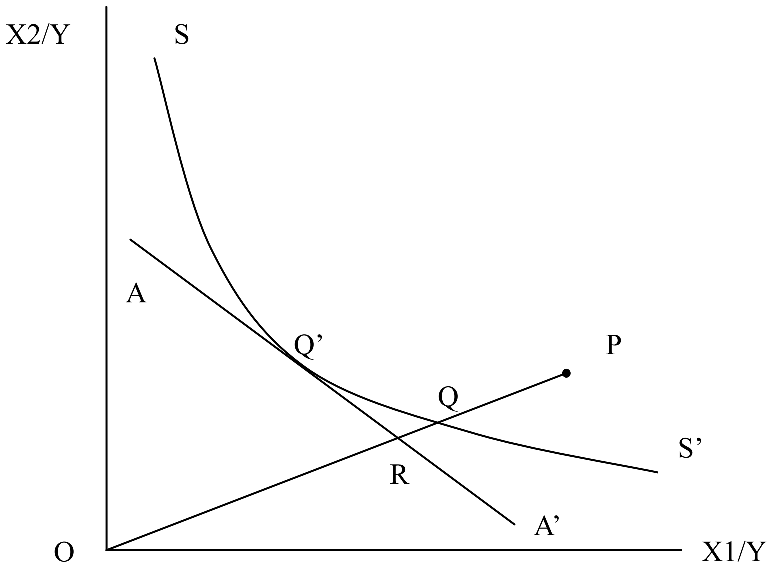

Figure 2.5: Farrell (1957)'s measure of technical and allocative efficiency

If firm P is productively efficient, both technically and allocatively efficient, its costs should be at the fraction OR/OP of what it actually is. This ratio is termed by Farrell as overall efficiency of the firm measured by product of technical and allocative efficiency: OE= TE*AE= OQ/OP*OR/OQ=OR/OP

The key contribution of Farrell is that he built up an efficient frontier termed as the unit isoquant SS’ as a benchmark to measure the relative performance of productive units. This underlying idea has been widely spread by other researchers from then, known as frontier analysis. Different techniques have been utilised to either calculate or estimate these efficient frontiers. These different techniques can be classified in different ways. The criterion followed here is to distinguish between parametric and non-parametric methods that is, between techniques where efficiency level is estimated from a pre-defined the functional form of the efficient frontier and those where no functional form is pre-established but efficiency level is calculated from the sample observations.

2.2.3 Non-parametric approaches

The non-parametric approach has been traditionally assimilated into data envelopment analysis (DEA). The aim of this method is to define a frontier envelopment surface for all the sample observations. The DEA frontier is formed as the piecewise linear combinations that connect the set of best-practice observations, yielding a convex production possibility set. Those units that lie on the frontier are the efficient firms. On the other hand, units that do not lie on that surface can be considered as inefficient and an individual inefficiency score will be calculated for each one of them.

A key drawback of DEA is that it does not take into account the impact of random error. It is assumed to be: (a) no measurement error in constructing the frontier; (b) no luck that temporarily gives a production unit better performance; (c) no inaccuracies created by accounting rules that would make measured outputs and inputs deviate from economic outputs and inputs; and (d) no sampling error caused by selection of a sample instead of conducting a census of the population. Any of these errors that appear in an inefficient unit’s data may be reflected as part of inefficiency. What could be more problematical is that any of these errors in one of the units that lie on the efficient frontier may alter the measured efficiency of all the units that are compared to this unit or linear combinations involving this unit.

2.2.4 Parametric approaches

Instead of calculating the efficiency from sample observations, parametric approaches aim to construct the efficient frontier by setting up econometric models to estimate the efficiency score. If one has input and output data in the banking sector, production function and production frontier can be constructed and technical efficiency can be measured as the deviation of a bank's actual output from the estimated production frontier at a given level of input. Alternatively, if we can have further information on input prices, the dual of production function, cost function can be used to measure cost efficiency. Cost efficiency is simply the extent to which a bank's cost is above the 'best' performing bank determined by the cost frontier for a given level of output in the same condition. According to different ways to specify this technical efficiency term, parametric approaches can be divided into deterministic frontier approach and stochastic frontier approach. The rest of this subsection is organised as follows. Deterministic frontier approach will be discussed briefly in section 2.2.4.1. Stochastic frontier approach, which is the most popular and adopted parametric approach in the modern literature, is reviewed extensively in section 2.2.4.2.

2.2.4.1 Deterministic Frontier Approach

Suppose firm i (i=1,…,N) is producing the scalar output y_i from a vector of M inputs, $\mathbf{x}_i$. The production technology is represented by $f(\mathbf{x}_i;\boldsymbol{\beta})$ as seen in Figure 2.6. If the producer lies on the frontier, the technical efficiency will equal to one. If not, say less than one, it will provide a shortfall between the observed output and the maximum feasible output. Under deterministic frontier approach, the entire shortfall of the observed output and the maximum feasible output is attributed to technical inefficiency. Therefore, the output oriented technical efficiency, TE_i, is defined by the ratio of the actual output level to the maximum level of output

$$TE_i = \frac{y_i}{f(\mathbf{x}_i;\boldsymbol{\beta})} \qquad [2.12]$$

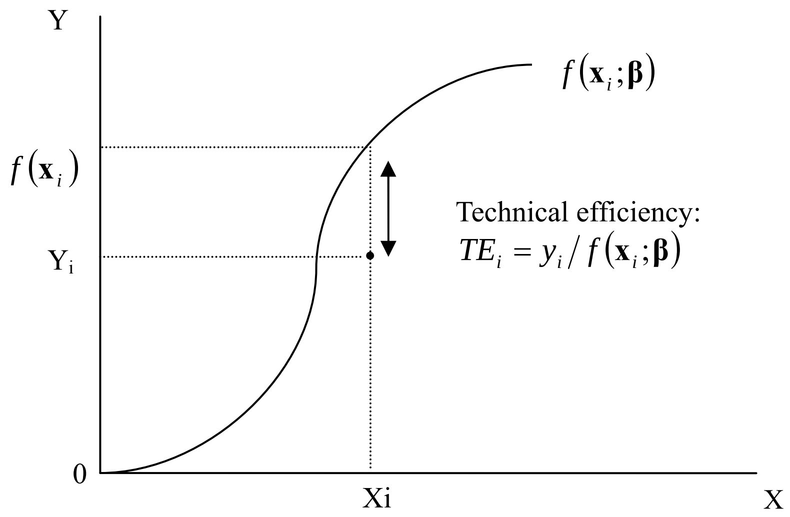

Figure 2.6: Production technology and technical efficiency (output oriented)#

Assume $TE_i = \exp(-u_i)$, [2.12] can be rewritten as:

$$y_i = f(\mathbf{x}_i;\boldsymbol{\beta}) \cdot \exp(-u_i) \qquad [2.13]$$

Since $TE_i = \exp(-u_i)$, we will have $\ln TE_i = -u_i$ which leads to $u_i = -\ln TE_i \approx 1 - TE_i$. Therefore, u_i can approximately measure the firm's technical inefficiency and represents the shortfall from the frontier for each firm. By adopting logarithm in both sides, the deterministic production frontier model becomes

$$\ln y_i = \ln f(\mathbf{x}_i;\boldsymbol{\beta}) - u_i \qquad [2.14]$$

Once the production technology is parameterized, both goal programming and econometric techniques like corrected ordinary least square (COLS) or modified ordinary least square (MOLS) can be applied to either calculate or estimate the parameter vector and to obtain the estimates of u_i. Then estimates of firm specific technical efficiency can be obtained by $TE_i = \exp(-u_i)$.

The drawback of deterministic frontier approach lies in that it treats the entire deviation of observed output and maximum feasible output as technical inefficiency. It assumes that all deviations from the efficient frontier are under the control of the firm. However, there are some circumstances out of the agent's control that can also determine firm's performance. Regulatory-competitive environments, weather, luck, socio-economic and demographic factors, uncertainty, etc. should not be considered as technical inefficiency. Moreover, any specification errors are also considered as inefficiency from the point of view of deterministic techniques. Therefore, a model is required to attribute those deviations from the efficient frontier to some combination of random errors and technical inefficiency. This model is developed and known as stochastic frontier approach. Since stochastic frontier approach models both specification failures and uncontrollable factors independent of the technical inefficiency component by introducing a two-sided random error into the specification of the frontier model, it now becomes the most popular and widely used parametric approach for efficiency measurement.

2.2.4.2 Stochastic Frontier Approach

2.2.4.2.1 Cross-sectional framework

Aigner *et al.* (1977), Meeusen and van den Broeck (1977) and Battese and Corra (1977) simultaneously introduced stochastic production frontier models. These models allow for technical inefficiency, but they also acknowledge the impact of random shocks outside the control of firms on their production, by including an idiosyncratic error term. Therefore, unlike in the deterministic frontier approach, the production frontier here is represented by $f(\mathbf{x}_i;\boldsymbol{\beta})\exp(v_i)$ as displayed in Figure 2.7.

Due to these possible random influences, the production frontier can vary across firms and over time across the same firm. On this interpretation, the frontier is stochastic, with random disturbance v_i greater or less than zero being the result of favourable or unfavourable external events such as luck, weather and sampling and measurement errors. Under the stochastic frontier, the firm will be considered as technically efficient if it lies on the frontier $f(\mathbf{x}_i;\boldsymbol{\beta})\exp(v_i)$. Any deviation will be termed as technical inefficiency, measured by the one-sided error term u_i as the result of factors under the firm's control. As seen in Figure 2.7, it is measured as the overall shortfall of the observed output from the frontier. And technical efficiency will be written as

$$TE_i = \frac{y_i}{f(\mathbf{x}_i;\boldsymbol{\beta})\exp(v_i)} \qquad [2.15]$$

Recall $TE_i = \exp(-u_i)$, firms' actual output can be written as $y_i = f(\mathbf{x}_i;\boldsymbol{\beta})\exp(v_i - u_i)$.

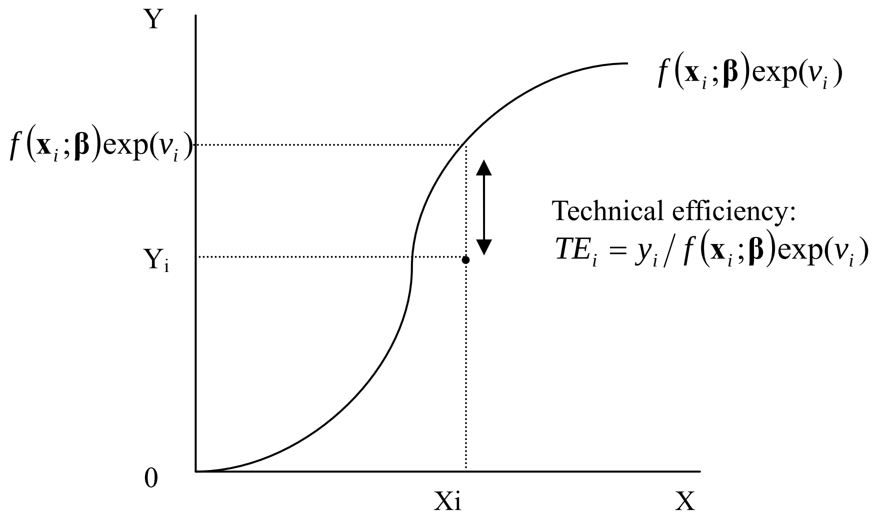

Figure 2.7: Production technology and technical efficiency

By taking the log terms, the general stochastic production frontier model can be written as follows.

$$\begin{aligned} \ln y_i &= \ln f(\mathbf{x}_i;\boldsymbol{\beta}) + \varepsilon_i \\ \varepsilon_i &= v_i - u_i \end{aligned} \qquad [2.16]$$

where ε_i is a composed error term consisting of two components v_i and u_i. v_i represents the two-sided noise component and u_i is the nonnegative technical inefficiency term. The noise component v_i is assumed to be independently, identically distributed and symmetrically distributed independently of u_i.

Estimation of technical efficiency relies on the choice of functional form and estimates of parameters $\boldsymbol{\beta}$. Cobb-Douglas functional form is the most popular choice in 1980s but replaced by the use of more flexible translog functional form. While in banking studies, as surveyed in the last section, the Fourier flexible function can also be used if global approximations are desirable at the expense of more complex parameterization and possible shortage of degrees of freedom in the estimation. To estimate parameters $\boldsymbol{\beta}$, one can use either ordinary least square estimation (OLS) or the maximum likelihood estimation (MLE). Under the assumption that u_i are distributed independently of the inputs, OLS provides consistent estimates of all the parameters except the intercept. However, to estimate technical inefficiency of each producer, a consistent estimate of the intercept is required. Therefore a two-step procedure is utilized here in which in the first step, OLS is used to estimate the slope parameters, and the second step involves the use of MLE to estimate the intercept and the variance of the two error components. In the estimation of technical inefficiency of each producer, one needs to separate the estimates of statistical noise v_i and technical inefficiency term u_i. In other words, technical inefficiency term u_i needs to be extracted from the estimates of ε_i, and this requires the distributional assumption on the two error components.

The following one-sided distributions have been employed: the half-normal and exponential distributions proposed by Aigner *et al.* (1977), the truncated normal proposed by Stevenson (1980)[1], and the two-parameter Gamma distribution proposed by Greene (1990). Other composed-error distributions could be constructed following Greene's (1990) methodology. In this short introduction of the stochastic frontier analysis I concentrate on the normal and half-normal case due to their popularity in empirical applications.

Suppose a stochastic frontier model given in [2.16] has the following distributional assumptions:

i) $v_i \sim \text{iid } N\left(0, \sigma_v^2\right)$;

ii) $u_i \sim \text{iid } N^+\left(0, \sigma_u^2\right)$, that is, a nonnegative half normal;

iii) v_i and u_i are distributed independently of each other and of the regressors.

[1] Derivations of the density functions of these three distributions are provided in the **Appendix 1** along with the mean and variance of the inefficiency term. In my knowledge, these derivations have not been provided in any previous literature.

The density function of v is

$$f(v) = \frac{1}{\sqrt{2\pi}\sigma_v}.\exp\left\{-\frac{v^2}{2\sigma_v^2}\right\} \qquad [2.17]$$

and the density function of u is given by the function

$$f(u) = \frac{2}{\sqrt{2\pi}\sigma_u}.\exp\left\{-\frac{u^2}{2\sigma_u^2}\right\} \qquad [2.18]$$

Given the independence assumption, the joint density function of v_i and u_i is the product of their individual density functions, and so

$$f(u,v) = \frac{2}{2\pi\sigma_u\sigma_v} \cdot \exp\left(-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right) \qquad [2.19]$$

Since $\varepsilon_i = v_i - u_i$, the joint density functions for u_i and ε_i is

$$f(u,v) = \frac{2}{2\pi\sigma_u\sigma_v} \cdot \exp\left(-\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon+u)^2}{2\sigma_v^2}\right) \qquad [2.20]$$

The marginal density function[2] of ε_i is obtained by integrating u out of $f(u,\varepsilon)$, which yields

$$f(\varepsilon) = \frac{2}{\sigma} \cdot \phi\left(\frac{\varepsilon}{\sigma}\right) \cdot \left[1 - \Phi\left(\frac{\varepsilon\lambda}{\sigma}\right)\right] \qquad [2.21]$$

where $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}, \lambda = \sigma_u / \sigma_v$. The parameterizations of σ, λ are quite convenient because '*λ is thereby interpreted to be an indicator of the relative variability of the two sources of random error that distinguish firms from one another*' (Aigner *et al.*, 1977). $\lambda^2 \to 0$ implies $\sigma_v{}^2 \to \infty$ and/or $\sigma_u{}^2 \to 0$, while the symmetric error term dominates the one-sided error component in the determination of ε_i. As $\lambda^2 \to \infty$ implies either $\sigma_v{}^2 \to 0$ or $\sigma_u{}^2 \to \infty$, while the one-sided error component dominates the symmetric error component in the determination of ε_i. The former case denotes to an OLS production function model without technical inefficiency, while the latter one indicates a deterministic production function model without random noise terms.

[2] This result is first introduced in Aigner *et al.* (1977) and it was taken as granted by other researchers using this technology in the banking efficiency study. So far, no derivation of this density function of composed error term has been seen in the literature. Therefore, I think it would be a good addition to include this derivation in my PhD thesis. Derivation is provided in **Appendix 2**.

By assuming the specific distribution for the composed error term, one can use the MLE to obtain the estimates for parameters $(\alpha, \beta, \sigma, \lambda)$. After obtaining the estimates of those parameters, one can obtain the estimates of technical inefficiency of each producer. Estimates of the composed error term of $\varepsilon_i = v_i - u_i$ contain information on u_i. $\varepsilon_i > 0$ indicates that u_i is not large (since the mean of noise is zero), suggesting that this producer is relatively efficient; whereas if $\varepsilon_i < 0$, u_i will be large, indicating that this producer is relatively inefficient. Therefore, to extract the information of u_i out of ε_i, one can use the conditional distribution of u_i given ε_i as it contains whatever information concerning u_i in ε_i. Based on this idea, two widely used estimators can be found in the modern literature, which are proposed by Jondrow *et al.* (1982) and Battese and Coelli (1988).

Jondrow *et al.* (JLMS) (1982) show that if $u_i \sim \text{iid } N^+(0, \sigma_u^2)$, the conditional distribution of u_i given ε_i was

$$f(u \mid \varepsilon) = \frac{f(u, \varepsilon)}{f(\varepsilon)} = \frac{1}{\sqrt{2\pi}\sigma_*} \cdot \exp\left\{-\frac{(u - \mu_*)^2}{2\sigma_*^2}\right\} \Big/ \left[1 - \Phi\left(-\frac{\mu_*}{\sigma_*}\right)\right] \quad [2.22]$$

where $\mu_* = -\varepsilon\sigma_u^2 / \sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2 / \sigma^2$. They argue that since the conditional distribution of u_i given ε_i is $N^+(\mu_*, \sigma_*^2)$, to obtain a point estimator of u_i, one can use either the mean or the mode of its conditional distribution. They are given by:

$$E(u_i \mid \varepsilon_i) = \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i} / \sigma_*)}{1 - \Phi(-\mu_{*i} / \sigma_*)}\right] = \sigma_*\left[\frac{\phi(\varepsilon_i\lambda / \sigma_*)}{1 - \Phi(\varepsilon_i\lambda / \sigma_*)} - \left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right] \quad [2.23]$$

and
$$M(u_i \mid \varepsilon_i) = \begin{cases} -\varepsilon_i\left(\dfrac{\sigma_u^2}{\sigma^2}\right) & \text{if } \varepsilon_i \le 0 \\ 0 & \textit{otherwise} \end{cases} \quad [2.24]$$

Once the point estimates of u_i are obtained, estimates of technical efficiency of each producer can be obtained from $TE_i = \exp(-\hat{u}_i)$, where $\hat{u}_i$ is either $E(u_i \mid \varepsilon_i)$ or $M(u_i \mid \varepsilon_i)$.

Battese and Coelli (B&C) (1988) proposed the alternative point estimator for TE_i:

$$TE_i = E\left(\exp\{-u_i\} \mid \varepsilon_i\right) = \left[\frac{1-\Phi(\sigma_* - \mu_* / \sigma_*)}{1-\Phi(-\mu_* / \sigma_*)}\right] \cdot \exp\left\{-\mu_* + \frac{1}{2}\sigma_*^2\right\}$$

[2.25]

Their estimator is preferred to the JLMS estimators because $1 - E(u_i \mid \varepsilon_i)$ only includes the first order term in the approximation of power series $\exp(-u_i \mid \varepsilon_i)$. Therefore, the B&C estimator can be viewed as the exact expression of the mean of the distribution of technical efficiency, whereas the JLMS estimates are the exact expressions of the central tendencies of a first order approximation to the distribution of technical efficiency.

Unfortunately, all these estimators suffer an important defect, namely that although they are unbiased they are not consistent estimators of u_i, since regardless of N, the variance of the estimate remains nonzero (see Greene, 1993: 81). Also, Greene (1993) argued that "*this inconsistency of the estimator of u_i is unfortunate in view of the fact that the purpose of the exercise to begin with is to estimate inefficiency. It would appear, however, that no improvement on this measure for the single-equation, cross-sectional framework considered here is forthcoming.*"

Besides this inconsistency of the estimator of inefficiency, there are two other difficulties with the cross-sectional stochastic production frontier models. First, in all cross-sectional studies, strong distributional assumptions must be imposed on the error term. With different distributional assumptions over the one-sided error term, although the relative ranking of firms based on inefficiency calculations seems unaffected, the absolute level of inefficiencies differs. Second, when using MLE to estimate the parameters and inefficiency, one needs to assume that technical inefficiency is independent of the regressors, although it is easy to imagine that technical inefficiency might be correlated with the input vectors producers selected. However, these three limitations associated with the cross-sectional stochastic frontier models can be avoided if panel data is available.

2.2.4.2.2 Panel data framework

The above three limitations associated with the cross-sectional stochastic frontier model were first noted in Schmidt and Sickles (1984). However, each of these limitations is avoidable if one has access to panel data. First, conventional panel data estimation techniques can be used to measure the technical efficiency if the panel data is available, and not all of these techniques rest on the strong distributional assumption. Repeated observations on a sample of producers can serve as a substitute for strong distributional assumptions. Second, not all the panel data estimation techniques require the independence of technical inefficiency from the regressors. Repeated observations of producers can also serve as a substitute for the independence assumption. Since adding more observations on each producer generates information not provided by adding more producers to a cross section, technical efficiency of each producer in the sample can be measured consistently as $T \to \infty$. Repeated observations on a sample of producers resolve the inconsistency problem with the JLMS technique. However, this final benefit of using panel data techniques can be overstated, however, since many panels are relatively short.

In regard to panel data production frontier models, those models in which technical efficiency is allowed to vary across producers but is assumed to be time-invariant for each producer will be discussed first. However, assuming technical efficiency to be time invariant can be tenuous, especially in long panels. Therefore, models which allow technical efficiency not only to vary across producers but also throughout time for each producer will be considered, known as time varying technical efficiency. All the panel data models in the next two subsections are assumed to be balanced panel, measuring that each producer is observed T times. However, unbalanced panels, in which producer i can be observed T_i times, which could be less than T, can be easily accommodated into all the panel data models discussed here. My discussion draws on Kumbhakar and Lovell (2000).

2.2.4.2.2.1 Time-invariant technical efficiency

Suppose firm i (i=1,…,N) that produces a scalar output from M inputs are observed for

T times. Cobb-Douglas production technology is assumed for simplicity. Therefore, the production frontier with time-invariant technical efficiency can be written as

$$\ln y_{it} = \beta_{ot} + \sum_m \beta_m \ln x_{mit} + v_{it} - u_i \qquad i = 1 \ldots N; t = 1 \ldots T$$

[2.26]

where v_{it} represents random statistical noise while $u_i \geq 0$ represents technical inefficiency. No technical change is allowed in this model. For T=1, this model is exactly the one suggested by Aigner *et al.* (1977). For T>1, it fits the conventional panel data model with producer effects but without time effects. The only difference is that the producer effects are assumed to be one sided (nonnegative) as technical inefficiency.

2.2.4.2.2.1.1 Fixed-effects model

The simplest panel data model is the fixed-effects model since it only has a few assumptions in the error terms. v_{it} are assumed to be iid$(0, \sigma_v^2)$ and are uncorrelated with the regressors. No distributional assumption is made in u_i, and therefore, u_i is allowed to be correlated with the regressors or with the random noise v_{it}. u_i is treated as fixed (i.e., nonrandom) effects that become producer specific intercept parameters to be estimated along with the $\beta_m s$ using OLS. The model becomes

$$\ln y_{it} = \beta_{0i} + \sum_m \beta_m \ln x_{mit} + v_{it}$$

[2.27]

where $\beta_{0i} = \beta_0 - u_i$ is the producer specific intercept. Least square with dummy variable technique (LSDV for short) can be used to estimate the producer specific intercept. It can be done in three equivalent ways: 1) by suppressing the constant term and adding a dummy variable for each of the N producers, or 2) by keeping the constant term and adding (N-1) dummies, or 3) using the within transformation, in which all the data will be expressed in terms of deviation from producer means (e.g., replace y_{it} by $y_{it} - \bar{y}_i$, etc) and the N intercepts are recovered as means of the firm specific residuals by firm.

When N estimated intercepts $\hat{\beta}_{01},\dots,\hat{\beta}_{0n}$ are obtained, simply define that

$$\hat{\beta}_0 = \max_i(\hat{\beta}_{0i}) \qquad [2.28]$$

and then technical inefficiency u_i can be estimated from

$$\hat{u}_i = \hat{\beta}_0 - \hat{\beta}_{0i} \qquad [2.29]$$

which ensures that all $\hat{u} \geq 0$. The producer specific estimate of technical efficiency are then given by

$$TE_i = \exp\{-\hat{u}_i\} \qquad [2.30]$$

Same as in the COLS model under cross-sectional data, in fixed-effects model at least one producer is assumed to be 100% technically efficient and technical efficiency of other firms is measured relative to that efficient producer.

The fixed-effects model has nice consistency properties, the LSDV estimates of the β_{0i} are consistent as either N or $T \to \infty$. Consistency of the individual LSDV estimate of intercept β_{0i}, however, requires $T \to \infty$. The LSDV estimates of the producer specific inefficiency term u_i requires both N and $T \to \infty$. Neither of these consistency properties requires the assumption of distribution of the random noises v_{it}. In contrast to the MLE cross-sectional model, the fixed-effects panel data model provides the consistent estimates of producer specific technical efficiency.

However, the fixed-effects model has a potential serious drawback. By using the within transformation, it is possible to include in the specification regressors that are time invariant but varying across firms. Therefore, the estimated fixed effects u_i, which is intended to capture the producer specific technical inefficiency, will unfortunately capture some fixed effects of variables (e.g. regulatory environment, capital shock, etc) that are not in any sense representing technical inefficiency. This confusion may occur whether or not other effects are included as regressors in the model. To avoid this problem, one must make assumptions about the uncorrelatedness of effects and regressors and/or about the distribution of the effects, which leads us to the random-effects panel data model.

2.2.4.2.2.1.2 Random-effects Model

In the fixed-effects model, u_i is treated as fixed but can be correlated with the regressors. Now, considering an opposite situation where u_i is assumed to be randomly distributed with constant mean and variance but they are uncorrelated with the regressors. This modification allows one to include some time invariant variable in the model. However, at this point, we still make no assumptions about the distribution of the effects, although they are still assumed to be nonnegative. The assumption of the random noise v_{it} is as before. With these modifications of assumptions, [2.26] can be rewritten as

$$\begin{aligned}\ln y_{it} &= [\beta_0 - E(u_i)] + \sum_m \beta_m \ln x_{mit} + v_{it} - [u_i - E(u_i)] \\ &= \beta_0^* + \sum_m \beta_m \ln x_{mit} + v_{it} - u_i^*\end{aligned}$$

[2.31]

Then random-effects producer specific technical efficiency can be estimated either by using the two-step generalised least square (GLS) method, or by maximum likelihood estimation (MLE) method if further distributional assumption on the error components is tenable (e.g. normal and half normal distributional assumption used in Pitt and Lee, 1981 and normal and truncated normal distributional assumption used in Kumbhakar, 1987 and Battese and Coelli, 1988).

In the case of no distributional assumption made on the error component, GLS is the appropriate means of estimating the producer specific technical efficiency. In the first step, OLS is used to obtain parameters estimates. In the second step β_0^* and the $\beta_m s$ are then reestimated using feasible GLS. Notice that β_0^* does not depend on *i*, since $E(u_i)$ is a positive constant, so there is only one intercept term to be estimated. Once β_0^* and the $\beta_m s$ have been estimated using feasible GLS, the u_i^* can be estimated by means of the residuals for firm i:

$$\hat{u}_i^* = \frac{1}{T} \sum_t \left(\ln y_{it} - \hat{\beta}_0^* - \sum_m \hat{\beta}_m \ln x_{mit} \right)$$

[2.32]

The estimator of u_i is obtained by means of the normalization

$$\hat{u}_i = \max_i(\hat{u}_i^*) - \hat{u}_i^* \qquad [2.33]$$

An alternative estimator of u_i^* is the best linear unbiased predictor (BLUP). The BLUP of u_i^* is

$$\hat{u}_i^* = -\left[\frac{\hat{\sigma}_u^2}{T\hat{\sigma}_u^2 + \hat{\sigma}_v^2}\right] \cdot \sum_t \left(\ln y_{it} - \hat{\beta}_0^* - \sum_m \hat{\beta}_m \ln x_{mit} \right) \qquad [2.34]$$

Both estimates from the above two alternative methods are consistent as both N and $T \to \infty$. Estimates of the producer specific technical efficiency are then obtained by substituting the estimated $\hat{u}_i$ into $TE_i = \exp\{-\hat{u}_i\}$. As in the fixed-effects model, the estimator of technical efficiency in the random-effects model requires that at least one producer is 100% technically efficient, with technical efficiency of the remaining producers being measured relative to the technically efficient producers.

The biggest advantage of the random-effects model lies in the random settlement of u_i, which allows time-invariant regressors in the model. Therefore, these time-invariant effects would not contaminate the measurement of technical efficiency. Consequently, an increased level of efficiency score can be observed. However, it has to sacrifice the freedom of allowing u_i to be correlated with the regressors presumed in the fixed-effects model. Hausman and Taylor (1981) have developed an uncorrelatedness test to test the significance of differences between the fixed-effects estimator and the GLS estimator.

So far, no distributional assumption is made on u_i, demonstrating the capacity of having access to the panel data to avoid either the strong distributional assumption or the equally strong independence assumption usually made in the conventional cross sectional production frontier literature. However, if such distributional assumption is tenable in the panel data context, MLE can be used to estimate the time-invariant producer specific technical efficiency.

Pit and Lee (1981) made the following normal and half-normal distributional assumption on the error components in panel data stochastic production frontier model and obtained the maximum likelihood estimates for a sample of Indonesian weaving firms. The distributional assumptions are:

(i) $v_{it} \sim$ iid $N\left(0, \sigma_v^2\right)$

(ii) $u_i \sim$ iid $N^+\left(0, \sigma_u^2\right)$

(iii) u_i and v_i are distributed independently of each other, and of the regressors.

Then technical efficiency can be estimated by using the MLE in the same manner as in the cross-sectional framework. First, one needs to derive the log likelihood function for the composed error term and get the estimates for $\left(\beta_0, \beta_m, \sigma_u^2, \sigma_v^2\right)$. Since u_i is independent of time, the density function of u_i remains the same as in the cross sectional framework, given by $f(u) = \frac{2}{\sqrt{2\pi}\sigma_u} \cdot \exp\left\{\frac{-u^2}{2\sigma_u^2}\right\}$. However, the density function of **v**, denoted as a vector of $v_{\text{it}} = \left(v_{i1}, \ldots, v_{iT}\right)'$, is given by the following generalization of density function of v_i in the cross-sectional framework, which is now

$$f(\mathbf{v}) = \frac{1}{(2\pi)^{T/2}\sigma_v^T} \cdot \exp\left(\frac{-\mathbf{v}'\mathbf{v}}{2\sigma_v^2}\right) \qquad [2.35]$$

Given the independence assumption, the joint density function of u_i and **v** is

$$f(u, \mathbf{v}) = \frac{1}{(2\pi)^{(T+1)/2}\sigma_u \sigma_v^T} \cdot \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{\mathbf{v}'\mathbf{v}}{2\sigma_v^2}\right\} \qquad [2.36]$$

and the joint density function of u_i and ε is

$$f(u, \varepsilon) = \frac{1}{(2\pi)^{(T+1)/2}\sigma_u \sigma_v^T} \cdot \exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2} - \frac{\varepsilon'\varepsilon}{2\sigma_v^2} + \frac{\mu_*^2}{2\sigma_*^2}\right\} \qquad [2.37]$$

where $\mu_* = -\frac{T\sigma_u^2\bar{\varepsilon}}{\sigma_v^2 + T\sigma_u^2}$, $\sigma_*^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma_v^2 + T\sigma_u^2}$, $\bar{\varepsilon} = \frac{1}{T}\sum_t \varepsilon_{it}$

Thus the marginal density function of ε is

$$f(\varepsilon) = \int_0^\infty f(u, \varepsilon)du = \frac{2\left[1 - \Phi\left(-\mu_*/\sigma_*\right)\right]}{(2\pi)^{T/2}\sigma_v^{T-1}\left(\sigma_v^2 + T\sigma_u^2\right)^{1/2}} \cdot \exp\left(-\frac{\varepsilon'\varepsilon}{2\sigma_v^2} + \frac{\mu_*^2}{2\sigma_*^2}\right)$$

[2.38]

The log likelihood function for a sample of N producers, each observed for T times, is

$$\ln L = \text{constant} - \frac{\mathrm{N(T-1)}}{2}\ln\sigma_v^2 - \frac{N}{2}\ln\left(\sigma_v^2 + T\sigma_u^2\right) + \sum_i \ln\left[1-\Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)\right]$$
$$-\frac{\sum_i \varepsilon_i^{'}\varepsilon_i}{2\sigma_v^2} + \frac{1}{2}\sum_i\left(\frac{\mu_{*i}}{\sigma_*}\right)^2$$

[2.39]

After having the log likelihood function, one can obtain the maximum likelihood estimates of $\beta, \sigma_v^2, \sigma_u^2$ and the producer specific time-invariant technical efficiency. To do so, first, one has to derive the conditional distribution $(u \mid \varepsilon)$

$$f(u \mid \varepsilon) = \frac{f(u,\varepsilon)}{f(\varepsilon)} = \frac{1}{(2\pi)^{1/2}\sigma_*\left[1-\Phi(-\mu_*/\sigma_*)\right]}\cdot\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}$$

[2.40]

which is exactly the density function of a variable distributed as $N^+\left(\mu_*, \sigma_*^2\right)$. Thus, either JLMS estimator or B&C estimator can be used as proxy of technical efficiency. For JLMS technical efficiency estimator, either the mean or the mode of the above distribution can be used as a consistent point estimator of technical efficiency as $T \to \infty$, given by

$$E(u_i \mid \varepsilon_i) = \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)}\right] \quad [2.41]$$

and
$$M(u_i \mid \varepsilon_i) = \begin{cases}\mu_{*i} & \text{if } \varepsilon_i \le 0\\ 0 & \text{otherwise}\end{cases} \quad [2.42]$$

Kumbhakar (1987) and Battese and Coelli (1988) use the Battese and Coelli efficiency calculation. The suggested log likelihood function for the normal-truncated normal case is given by

$$\ln L = \text{constant} - \frac{N(T-1)}{2}\ln\sigma_v^2 - \frac{N}{2}\ln(\sigma_v^2 + T\sigma_u^2) - N\ln\left[1-\Phi(-\frac{\mu}{\sigma_u})\right]$$
$$+\sum_i \ln\left[1-\Phi(-\frac{\tilde{\mu}_i}{\sigma_*})\right] - \frac{\sum_i \varepsilon_i\varepsilon_i}{2\sigma_v^2} - \frac{N}{2}\left(\frac{\mu}{\sigma_u}\right)^2 + \frac{1}{2}\sum_i\left(\frac{\tilde{\mu}_i}{\sigma_*}\right)^2$$

[2.43]

where $\tilde{\mu}_i = \dfrac{\mu\sigma_v^2 - T\bar{\varepsilon}\sigma_u^2}{\sigma_v^2 + T\sigma_u^2}$, $\sigma_*^2 = \dfrac{\sigma_u^2\sigma_v^2}{\sigma_v^2 + T\sigma_u^2}$, $\bar{\varepsilon} = \dfrac{1}{T}\sum_t \varepsilon_{it}$. Note that $\tilde{\mu}_i = \mu_{*i}$ if $\mu = 0$. The log likelihood function can be maximized with respect to the parameters so as to obtain maximum likelihood estimates of all parameters.

The conditional distribution $(u \mid \varepsilon)$ is given by

$$f(u \mid \varepsilon) = \frac{f(u,\varepsilon)}{f(\varepsilon)} = \frac{1}{(2\pi)^{1/2}\sigma_*[1-\Phi(-\tilde{\mu}/\sigma_*)]} \cdot \exp\left\{-\frac{(u-\tilde{\mu})}{2\sigma_*^2}\right\}$$

[2.44]

which is distributed as $N^+(\tilde{\mu}, \sigma_*^2)$. Either the mean or the mode of this distribution (JLMS) can serve as the basis of a point estimator of producer-specific time-invariant technical efficiency, given by

$$E(u_i \mid \varepsilon_i) = \tilde{\mu}_i + \sigma_*\left[\frac{\phi(-\tilde{\mu}_i/\sigma_*)}{1-\Phi(-\tilde{\mu}_i/\sigma_*)}\right] \qquad [2.45]$$

$$\text{and} \quad M(u_i \mid \varepsilon_i) = \begin{cases} \tilde{\mu}_i & \text{if } \tilde{\mu}_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad [2.46]$$

Alternatively Battese and Coelli estimator can be used

$$E(\exp\{-u_i\} \mid \varepsilon_i) = \frac{1-\Phi(\sigma_* - (\tilde{\mu}_i/\sigma_*))}{1-\Phi(-\tilde{\mu}_i/\sigma_*)} \cdot \exp\left\{-\tilde{\mu}_i + \frac{1}{2}\sigma_*^2\right\}$$

[2.47]

To summarize, three alternative panel data approaches are discussed: the fixed-effects model based on LSDV, the random-effects model based on FGLS, and the random-effects model based on MLE if the distributional assumption on inefficiency term is available. These three approaches have different requirements for the data, which generates researchers' preferences with regard to different issues. For instance, with large N and small T, or in the presence of time-invariant regressors, the random-effects approach is clearly preferred to the fixed-effects model. If independence of the effects and the regressors is a plausible assumption, MLE is generally more efficient than either LSDV or GLS, since it exploits distributional information which the other two do not. However, empirical evidences of using three approaches most or less report the similar efficiency ranking.

The advantage of the above panel data models is that one can choose from the testable assumptions that technical inefficiency u_i is fixed or random, or the ones that technical inefficiency u_i is uncorrelated with other regressors, or the additional distributional assumptions on v_i or u_i. However, these advantages come at the expense of a very strong assumption that technical inefficiency is constant over time. It would be an unrealistic assumption in many cases as it is expected that firms would respond differently to the new regulatory environment and the improvement of production technology. This response will have an impact on firms' overall performance and technical efficiency. Therefore, it is desirable to relax the assumption of time-invariant technical efficiency in the long time period and this desire leads to the development of time-varying panel data models in which technical efficiency is allowed to change over time.

2.2.4.2.2.2 Time-varying technical efficiency

Cornwell *et al.* (CSS) (1990) and Kumbhakar (1990) were among the first to propose a stochastic production frontier panel data model with time-varying technical efficiency. It modifies the panel data model in the context of time-invariant technical efficiency as seen in [2.26] by replacing u_i with u_{it} to allow the intercept and inefficiency to change over time.

$$\begin{aligned}\ln y_{it} &= \beta_{0t} + \sum_m \beta_m \ln x_{mit} + v_{it} - u_{it} \\ &= \beta_{it} + \sum_m \beta_m \ln x_{mit} + v_{it}\end{aligned} \qquad [2.48]$$

where β_{0t} is the production frontier intercept common to all producers in period t, $\beta_{it} = \beta_{0t} - u_{it}$ is the intercept for producer i in period t, and all other variables are as previously defined. Same as in the time-invariant panel data model, either fixed-effects or random-effects approach can be used to model the time-varying technical efficiency. If the distributional assumption is tenable, maximum likelihood approach can be pursued as well.

As in the time-invariant panel data models, the estimation of time-varying panel data models also involves two stages. In the first stage the objective is to estimate the parameters describing the structure of production technology, while in the second stage producer specific technical efficiency is obtained. However, with the current $\mathrm{N}\times\mathrm{T}$ panel setting, it is impossible to obtain estimates of all $\mathrm{N}\cdot\mathrm{T}$ intercepts β_{it}, the M slope parameters β_m, and σ_v^2. Therefore, to make this time-varying model applicable, further specifications of β_{it} or u_{it} is needed to reduce the number of parameters in the estimation.

In CSS paper, the authors used the following formulation to reduce the number of intercept parameters to $\mathrm{N}\cdot 3$.

$$\beta_{it} = \Omega_{i1} + \Omega_{i2}t + \Omega_{i3}t^2 \qquad [2.49]$$

where the intercept for producer i at time t is a quadratic function of t. Although this quadratic specification may still leave a lot of parameters to be estimated, especially if the ratio (N/T) is large, it provides a path to allow technical efficiency to vary both over producers and over time. Two special cases can also be examined associated with this specification. If $\Omega_{i2} = \Omega_{i3} = 0\ \forall \mathrm{i}$, this model will collapse to the time-invariant technical efficiency model described in [2.26]. If $\Omega_{i2} = \Omega_2\ \forall \mathrm{i}$ and $\Omega_{i3} = \Omega_3\ \forall \mathrm{i}$, this model collapses to a fixed-effects model with specific intercepts Ω_{i1} and a quadratic term of t common to all producers given by $\Omega_2 t + \Omega_3 t^2$. Two indistinguishable interpretations can be provided. On the one hand, it can be treated as technical efficiency is producer specific and varies through time in the same manner for each producer while on the other hand, it can be viewed as technical efficiency is producer specific and time-invariant but the quadratic time term captures the effect of technical change.

Analogous to time-invariant fixed- and random-effects model, this time-varying panel data model [2.48] and [2.49] can be estimated using within, GLS and efficiency instrumental variables (EIV) estimator, depending on whether technical efficiency is treated as fixed or random. For the fixed-effects approach, two estimation strategies can be used to obtain estimates of $(\Omega_{i1}, \Omega_{i2}, \Omega_{i3})$. The first attempt, suggested in CSS (1990), $(\Omega_{i1}, \Omega_{i2}, \Omega_{i3})$ can be estimated by omitting u_{it} from [2.49], estimating the $\beta_m s$ from residuals and then obtaining estimates of $(\Omega_{i1}, \Omega_{i2}, \Omega_{i3})$ for each producer from regressing the residuals on a constant, t and t^2. In the second procedure, as suggested by Kumbhakar and Lovell (2000), if N/T is relatively small, include u_{it} in [2.48], estimate Ω_{i1} as coefficient of producer dummies, and estimate $(\Omega_{i2}, \Omega_{i3})$ as the coefficients of producer dummies interacted with t and t^2. Once estimates of $(\Omega_{i1}, \Omega_{i2}, \Omega_{i3})$ are obtained, technical efficiency can be estimated by using the analogous procedure provided for time-invariant fixed-effects model, shown as:

$$\hat{\beta}_{it} = \hat{\Omega}_{i1} + \hat{\Omega}_{i2} t + \hat{\Omega}_{i3} t^2 \qquad [2.50]$$

$$\hat{\beta}_{0t} = \max_i \left(\hat{\beta}_{it}\right) \qquad [2.51]$$

$$\hat{u}_{it} = \hat{\beta}_{0t} - \hat{\beta}_{it} \qquad [2.52]$$

$$TE_{it} = \exp(-\hat{u}_{it}) \qquad [2.53]$$

As the time-invariant fixed-effects model cannot handle the potential existence of time-invariant regressors, time-varying fixed-effects model cannot as well. For this reason, CSS also produce a time-varying random-effects model to incorporate the time-invariant regressors. Estimates of technical efficiency can be obtained by using either GLS or EIV estimator. Same procedure shown in the fixed-effects approach applies for both GLS and EIV approaches but different sets of residuals are used.

Lee and Schmidt (1993) proposed an alternative specification for u_{it} in [2.48], which was specified as

$$u_{it} = \beta(t) \cdot u_i \qquad [2.54]$$

where $\beta(t)$ is specified as a function of a set of time dummy variables β_t. Time-varying technical efficiency can be estimated using both fixed- and random-effects models, in which β_t s are treated as coefficients of the (fixed or random) effects u_i. Once the β_t s and the u_i are estimated,

$$\hat{u}_{it} = \max_i \left(\hat{\beta}_t \hat{u}_i\right) - \hat{\beta}_t \hat{u}_i \qquad [2.55]$$

from which $TE_{it} = \exp(-\hat{u}_{it})$ can be calculated. The main advantage of this specification is that it allows technical efficiency to vary over time and it is more flexible than CSS model since it does not restrict u_{it} to any particular parametric term. However, the drawback of this approach lies in the fact that it is nonlinear and requires a more complicated estimator. As argued in Kumbhakar and Lovell (2000: 110), this model is appropriate for short panels, since it requires estimation of T-1 additional parameters (the β_t less one normalizing variable $\beta_1 = 1$).

Since then a number of similar specifications have been proposed for the time-varying technical efficiency models consisting of [2.48] and [2.54], mainly focusing on adding specifications of $\beta(t)$. Kumbhakar (1990) specified $\beta(t)$ as the following parametric function of time t:

$$\beta(t) = \left[1 + \exp\left(\gamma t + \delta t^2\right)\right]^{-1} \qquad [2.56]$$

where $\beta(t)$ has to be between zero and one and satisfies properties of monotonically increasing or decreasing, and concave or convex, depending on the signs and magnitudes of the two parameters γ and δ. The time-invariant random-effects model is a special case when $\gamma = \delta = 0$. The main advantage of Kumbhakar's specification is that it only requires two additional parameters to be estimated, γ and δ, compared to $N \cdot 3$ additional parameters in CSS model and T-1 additional parameters in Lee and Schmidt model. MLE techniques (see Kumbhakar and Lovell, 2000:110-112) are used by Kumbhakar to estimate the model consisting of [2.48], [2.54] and [2.56]. Therefore, additional independence and distributional assumptions on v_{it} and u_i are required, although they are the same as in Pitt and Lee (1981) time-invariant random-effects model.

An alternative parameterization was proposed by Battese and Coelli (1992), in which $\beta(t)$ was specified as

$$\beta(t) = \exp\left[-\eta(t - T)\right] \qquad [2.57]$$

where $\beta(t)$ satisfies the properties of (i) $\beta(t) \geq 0$ and (ii) $\beta(t)$ decreases at an increasing rate if $\eta > 0$ and increase at an increasing rate if $\eta < 0$, or remain the same if $\eta = 0$, which returns to the time-invariant random-effects model. Compared with Kumbhakar (1990) model, Battese and Coelli (1992) model also use MLE techniques to estimate the parameters and technical efficiency but it only has one additional parameter η to be estimated. However, both Kumbhakar (1990) specification [2.56] and Battese and Coelli (1992) formulation [2.57] suffer one drawback that technical efficiency has to vary in a monotonic pattern. Therefore, Battese and Coelli (1992) also propose an alternative specification of $\beta(t)$ to allow the nonmonotonic variation of technical efficiency. $\beta(t)$ is defined as

$$\beta(t) = 1 + \eta_1(t - T) + \eta_2(t - T)^2 \qquad [2.58]$$

where η_1 and η_2 are the two parameters to be estimated and once again, the time-invariant technical efficiency model is a special case if $\eta_1 = \eta_2 = 0$.

A limitation of the above time-invariant and time-varying panel data models is limited in the presence of unobserved time-invariant heterogeneities. In the time-invariant fixed- and random-effects model, u_i is intended to capture all but only time-invariant inefficiency. However, if there are time-invariant heterogeneities, they must appear in u_i whether they belong there or not. Worse yet, even some of the time-invariant heterogeneities can be identified, they cannot be modeled in the fixed-effects model by the virtue of LSDV fixed effects estimator. When analyzing the WHO panel on 191 countries, Greene (2004) argued that under either interpretation (fixed- or random-effects), the inefficiency term u_i will absorb a large amount of cross-country heterogeneities that would be measured inappropriately as inefficiency. In the time-varying fixed- and random-effects model, u_{it} is supposed to capture all but only time-invariant and time-varying inefficiency. But if any time-invariant heterogeneity exists, they will unfortunately be absorbed into u_{it}. Like the time-invariant fixed-effects model, the time-varying fixed-effects model cannot include any time-invariant heterogeneity due to LSDV estimator as well. Therefore, the above panel data models must be modified to address the presence of time-invariant heterogeneities.

The first attempt was proposed by Kumbhakar and Hjalmarsson (1993) in which the authors applied the same time-varying technical efficiency model as shown in [2.48] with u_{it} being broken down to two components: a producer-specific component capturing producer heterogeneity and a producer- and time-specific component representing technical efficiency. Therefore, u_{it} are specified as

$$u_{it} = \tau_i + \xi_{it} \qquad [2.59]$$

where τ_i captures the possible omitted time-invariant heterogeneities and ξ_{it} is the technical efficiency component assumed to follow the distribution $N^+\left(0, \sigma_\xi^2\right)$. Since $\xi_{it} > 0$, its parameter can be separately identified from that of v_{it}, which is assumed to be distributed as $N\left(0, \sigma_v^2\right)$. The estimation of this model involves two stages. In the first stage, either a fixed- or random-effects model can be used to estimate all parameters but $\beta_{0t} + \tau_i$ and the parameters associated with ξ_{it} and v_{it}. In the second stage, distributional assumptions on ξ_{it} and v_{it} are imposed and $\beta_{0t} + \tau_i$ and the parameters associated with ξ_{it} and v_{it} are estimated using conditional maximum likelihood estimation based on the parameters estimated in the first stage. The virtue of this approach is that it introduces τ_i to capture the time-invariant attributes and the distributional assumptions would not be imposed until the second stage of estimation. However, the main problem of this model also lies in the introduction of τ_i as it captures not only the time-invariant heterogeneities but also time-invariant technical efficiency, which is intended to be captured in the technical efficiency component ξ_{it}.

Greene (2004, 2005) explored the issue by reformulating the stochastic frontier specifically with introduction of the 'true', in his term, fixed-effects and random-effects model. The 'true' fixed-effects formulation is written as

$$\ln y_{it} = \beta_i + \sum_m \beta_m \ln x_{mit} + v_{it} - u_{it} \qquad [2.60]$$

where β_i is the firm specific intercept intended to capture all the time-invariant heterogeneities. It retains the distributional assumptions of stochastic frontier model but allows the inefficiency to vary freely over time. Regressors, inefficiency term and random error term are mutually uncorrelated but the firm's heterogeneity term is allowed to be correlated with the included variables. This model is estimated by MLE rather than the usual within group least square estimation employed in fixed-effects model for the reason that although within group least square approach can estimate unbiased and consistent $\beta_m s$, but it provides no measure of firm specific constant term β_i. However, this 'true' fixed-effects model has hardly been used in the literature (a recent exception can be found in Saal *et al.*, 2007) due to the incidental parameters problem and the concern that this model is over specified. The incidental parameters problem addresses the problem of a persistent bias in the estimates of the main parameters in the fixed-effects model especially when the sample period is relatively short. It has been widely discussed in binary choice models but not systematically tested in stochastic frontier models. Greene (2005) is amongst the first to discuss this problem and in his application to the US banking industry, the author find surprisingly small bias in the parameter estimates and more importantly only minor bias appeared to be transmitted to the estimates of technical inefficiency. The second and more vital problem is that this 'true' fixed-effects model may now overcompensate for the time-invariant heterogeneities since the firm specific constant term β_i, intended to capture all but only time-invariant heterogeneous factors will inevitably capture the time-invariant inefficiency. In other words, if there is persistent inefficiency, it will be absorbed into β_i rather than u_{it} where it belongs.

Besides this 'true' fixed-effects model, Greene (2004, 2005) also introduce a 'true' random-effects model. It is written as

$$\ln y_{it} = \beta_0 + \sum_m \beta_m \ln x_{mit} + v_{it} - u_{it} + \varpi_i \qquad [2.61]$$

where ϖ_i is a time-invariant, firm specific random term intended to capture cross firm time-invariant heterogeneities. This model has a precursor in the literature. Its specification is largely the same as that of Kumbhakar and Hjalmarsson (1993) in [2.48]. τ_i is meant to capture the perhaps omitted time-invariant inputs, just as ω_i here while ξ_{it} represents technical inefficiency as u_{it} does here. However, these two models differ in the term of estimation technique. Kumbhakar and Hjalmarsson (1993) use two-stage estimation strategy in which within group (LSDV) OLS or feasible GLS is used to estimate $\beta_m s$ followed by MLE of v_{it} and u_{it} with distributional assumption provided. While in Greene's 'true' random-effects model, MLE is used straightforward to estimate all the parameters. Detailed steps of MLE estimation are provided in Greene (2005:24-25). The ‘true’ random-effects model also has its shortcomings. As pointed out in Greene’s (2004) WHO panel study on 191 countries, he find that the ‘true’ random-effects model may become very unstable if any covariate are added into the model where the data are essentially lack of within group variation. It also shares the problem of overspecification as in the ‘true’ fixed-effects model. ϖ_i, meant to capture all but only time-invariant cross firm heterogeneities, will inevitably contract all the time-invariant technical inefficiency. Therefore, as the time-invariant and time-varying fixed- and random-effects model may overestimate the technical inefficiency in the presence of time-invariant heterogeneities, ‘true’ fixed- and random-effects models may underestimate it. This shared shortcoming of ‘true’ SFA models has been proved in both Greene’s (2004, 2005) applications of WHO and the US banking industries. Therefore, the “truth”, as argued in Greene (2004, 2005) is something between these models but nevertheless, it is better to address the time-invariant heterogeneities than to ignore it altogether as in [2.26].

2.2.5 Summary

This small subsection introduces the origin and recent development of frontier efficiency methods with particular interests in the parametric stochastic frontier approaches due to its capacity of considering the impact of random noise on firms’ performance. Compared with cross-sectional stochastic frontier framework, panel data stochastic frontier framework has the advantages of providing consistent estimates of parameters and inefficiency, possibility of avoiding the independence assumption of inefficiency and the regressors, and strong distributional assumptions of inefficiency (although proved in the literature that distributional assumptions are not so crucial in variations of inefficiency estimates, the model specification does. See Kumbakar and Lovell, 2000: 90 and Greene, 2008: 180-184).

In the panel data framework of the time-invariant and time-varying technical efficiency models, both the fixed- and random-effects approaches have virtues and shortcomings. The advantages of the time-invariant fixed-effects model are that it is distribution free, and it allows the inefficiency to be correlated with the regressors. However, this generality comes at the expense of measuring and identifying technical inefficiency against the relative "best" performed firms. This may result in a downward bias of efficiency estimates (see Kim and Schmidt, 2000). Moreover, time-invariant effects are treated ambiguously in the fixed-effects model. Although u_i is supposed to capture only the time-invariant inefficiency, it will also contain information of other time-invariant effects. The time-invariant random-effects model, such as Pitt and Lee (1981), allows the direct estimation of firm specific inefficiency term. However, it relies on the distributional assumption on inefficiency and the assumptions that u_i are uncorrelated with regressors in the model. Despite the above separate virtues, time-invariant fixed- and random-effects models share two shortcomings. They both assume technical inefficiency to be constant over time. However, in the long time period this would be a problematic assumption and it is more desirable to relax it and allow efficiency to vary over time. The existing literature contains several attempts in the manners of both fixed- and random-effects. CSS, Lee and Schmidt (1993), Kumbhakar (1990) and Battese and Coelli (1992) introduced different formulation on inefficiency to allow the variation of technical efficiency over time, but the variation of inefficiency has to obey certain structure and some are very rigid. The other shared weakness of these models is the lack of treatment of unobservable time-invariant heterogeneities. If any unobservable time-invariant heterogeneities exist, they will be forced to enter firm specific inefficiency term u_i. Even if some of these unobservable heterogeneities can be identified, they cannot be included in the fixed-effects model due to the within group least square estimation technique. Kumbhakar and Hjalmarrson (1993) and Greene (2004, 2005) tried to solve this problem by introducing a firm specific constant term to capture all the time-invariant heterogeneities. However, their parameterization may be overspecified since not only time-invariant heterogeneities will be removed from u_i but time-invariant inefficiency will also be pulled out. Therefore, a downward bias will be expected in u_i since it only measures time-varying inefficiency. As argued by Greene (2004, 2005), it is better to address the issue of heterogeneities than to ignore it altogether as in the fixed- and random-effects model. If some of the heterogeneities are identifiable, it is better to include them in the random-effects model to obtain more accurate estimates of technical inefficiency.

2.3 Measuring productivity change and its decomposition

2.3.1 Definition of Productivity

Productivity of a firm is usually defined as the ratio of the output(s) to the input(s). If there is only one output (Y) and one input (X), it boils down to the following expression:

$$\text{Productivity} = \text{output/input} = Y/X \qquad [2.62]$$

This formulation, however, is far too simple, because in the real world, most firms rely on more than one input in the production process. For example, in the banking industry, any commercial banks intend to use a combination of inputs, such as labour, deposits, fixed assets and others. Moreover, it is also more likely that more than one output is produced. For instance, a commercial bank provides loans as their main income source, but it also receives incomes from other earning assets, such as interbank bank loans, deposits in the central bank, and incomes from off-balance sheet business, like direct credit substitutes in which a bank substitutes its own credit for a third party; interest rate swaps; interest rate options; and currency options, and so on. Therefore, productivity discussed here usually refers to the multiple factor productivity or total factor productivity (TFP), which is measured involving all outputs in production, rather than other traditional partial productivity measures, such as labour productivity that measures output per worker. Although commonly used, such partial productivity measures can provide a misleading indication of overall performance of a firm.

To measure TFP, a method of aggregating the multiple inputs/outputs into a single input/output index has to be used to obtain a ratio measure of productivity. Suppose a firm produces J outputs from M inputs, where J and M are both larger than one, TFP can be measured as

$$TFP = \sum_{j=1}^{J} r_j y_j \Big/ \sum_{m=1}^{M} s_m x_m \qquad [2.63]$$

where $0 \le r_j \le 1, \sum_j r_j = 1$, $0 \le s_m \le 1, \sum_j s_m = 1$. r_j and s_m are weights for relative output and input. As well explained in Orea (2002), these weights have some useful properties, for instance, identity, monotonicity, separability and proportionality, which I will discuss later in this subsection.

2.3.2 Measuring Productivity

As early as Solow (1957), researchers have attempted to measure the productivity change as well as to identify its sources. Among these attempts, measurement of productivity can be categorized into two schools: non-frontier approaches and frontier approaches. It this subsection, a brief review of these two approaches will be provided.

2.3.2.1 Non-frontier approaches

This set of traditional approaches to productivity measurement generally assumes that observed output is the frontier output or best practice output. In this context, observed output is assumed to be technically efficient in the sense of Farrell (1957)[3]. Details of these approaches, as suggested by Grosskopf (1993), are overviewed in Diewert (1989) and Link (1987). By applying either non-parametric or parametric non-frontier approaches, the productivity change is simply measured as technical change. Discussions of non-frontier approaches are provided in Grosskopf (1993) and Coelli *et al.* (2005).

2.3.2.2 Frontier approaches

To illustrate the productivity measurement under frontier approaches, consider a production process that a firm uses a single input to produce a single output. As seen in Figure 2.8, the horizontal axis represents the single input x and the vertical axis represents the single output y. The line OF is the production frontier presenting the production technology at the moment. It shows the maximum output level that can be attained from the current given input level. Therefore, all the points lies on the frontier are terms as technically efficient, such as points B and C. Suppose the firm is operating at point A, then one can measure the productivity as the slope of the ray connecting the origin O and point A, which is y/x according to our definition of productivity defined at the beginning of this subsection. If the firm moves from point A to point B, it becomes technically efficient by producing more level of output with the same level of input. This improvement in technical efficiency also shows a higher slope of the ray OB than OA that indicates a high level of productivity for the firm. However, even though the firm is technically efficient when operating at point B, it can still improve its productivity by moving from point B to point C since the slope of the ray OC is higher than OB. The ray OC is the highest level the firm can achieve since it is tangent to the production frontier that represents the maximum possible productivity. This movement from point B to point C can be viewed as productivity gains from exploiting returns to scale. Here the firm is reducing its output size to move from decreasing returns to scale to constant returns to scale. Now consider another case and suppose the firm is initially producing at point A’. Its productivity can be improved by moving from A’ to the technically efficient frontier point B’. Besides this, the firm can also achieve higher

3 In contrast to Koopmans’s (1951) technical efficiency that is measured as the distance of the observed firm to the input isoquant, Farrell’s (1957) technical efficiency is measured as the distance of the observed firm to the input isoquant represented by the piecewise linear frontier.

productivity from exploiting returns to scale as well. However, to achieve that, the firm is increasing its output size to move from increasing returns to scale to constant returns to scale.

In the multiple outputs and multiple inputs case, if the input prices and output prices are available, two more measures can be contributed to the firm's productivity. One is from improvement in input mix allocative efficiency, in which a firm is trying to produce its outputs using the least-cost mix of inputs, given the information on input prices. For example, if the price of capital falls relative to the price of labour, the firm may be able to reduce its costs by using less labour and more capital. The other is from improvement in output mix allocative efficiency, in which a firm is trying to produce the optimal mix of outputs given the information on output price.

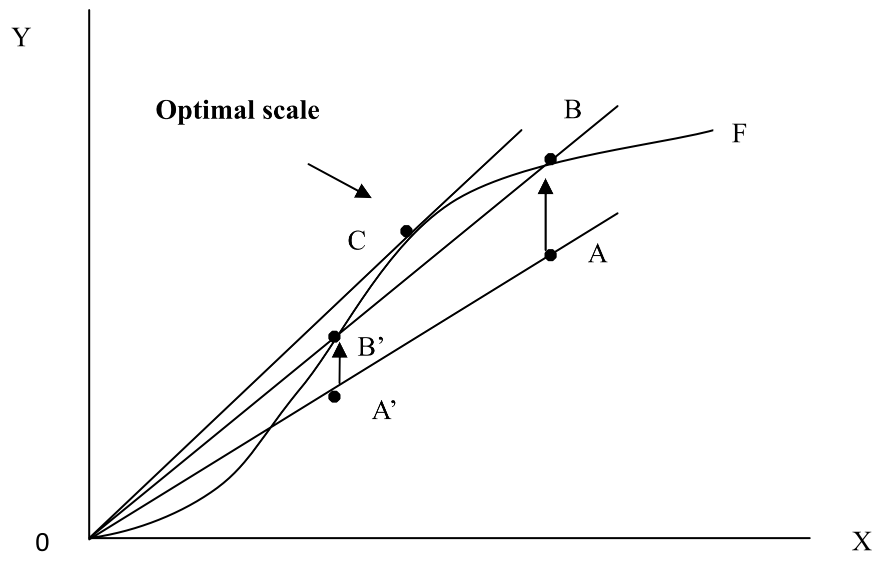

Figure 2.8: Production Frontier and Productivity

When comparing productivity over time, an additional productivity change, technical change, can be introduced (see Figure 2.9). Technical change measures the frontier shift over time. It captures the movement of production frontier F in period 0 to production frontier F' in period 1. Besides this technical change, technical efficiency change, scale economies change and input and output mix allocative efficiency change, if possible, can be measured as potential sources of TFP change.

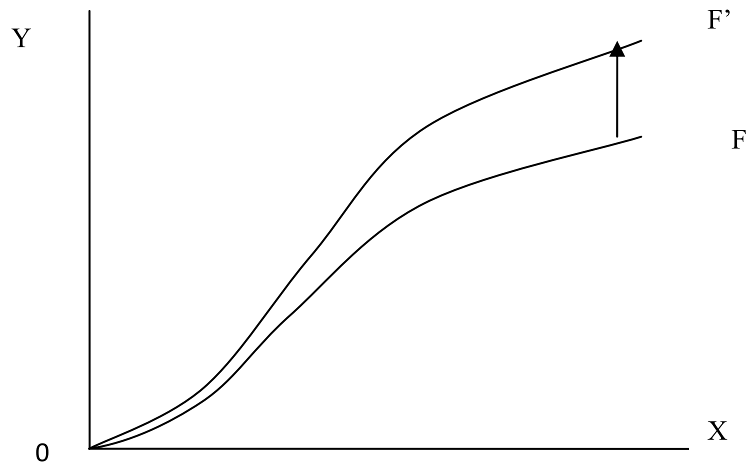

Figure 2.9: Technical change

Like non-frontier approaches, productivity measurement using frontier approaches can also be divided into categories of non-parametric frontier approaches and parametric frontier approaches.

2.3.2.2.1 Non-parametric frontier approaches

In the context of non-parametric approach, there are two ways to measure the productivity index. One is the partial oriented (either output oriented or input oriented) and the other is the simultaneously output and input oriented. The former was introduced by Caves *et al.* (1982) known as Malmquist productivity index calculated from the output or input distance functions and the latter was introduced by Bjurek (1996), termed as Malmquist total factor productivity index according to Lovell (2003). The large debate associated with these two approaches in the literature lies in whether researchers can provide economically meaningful decomposition of the respective productivity index. Caves *et al.* (1982) approach is far more popular than Bjurek (1996) approach for the reason that the former can be decomposed to several economically meaningful sources of the productivity growth but the decomposition issue has not been solved until Lovell (2003). Therefore, in this subsection, my focus will be on the Malmquist productivity index.

Constructed from distance function, Malmquist productivity index is calculated by using the non-parametric mathematical programming approach, DEA. The Malmquist productivity index is named after Malmquist (1953), who first proposed constructing input quantity indexes as ratios of distance functions. Although his objective is to compare alternative consumption bundles rather than actually measuring the productivity growth, his idea of constructing these indexes as ratios of distance function provides a natural way of modelling the production frontier as well as deviations from and shifts in that frontier, known as technical inefficiency change and technical change. This explains why Malmquist productivity index is actually introduced as a theoretical index by Caves *et al.* (1982) but named after Malmquist. In their paper, Caves, Christensen and Diewert defined the Malmquist productivity index, in the output oriented form, as the ratio of a pair of output distance functions and in the input oriented form as the ratio of a pair of input distance functions.

To define an output oriented Malmquist productivity index, assume that the firm use inputs $x \in \Re_{+}^{M}$ to produce outputs $y \in \Re_{+}^{J}$ with the production technology T that $\mathrm{T} = \{(\mathrm{y}, \mathrm{x}) : \mathrm{x} \text{ can producey}\}$ is a set of all technologically feasible output-input combinations. The output possibility set $\mathrm{P}(x) = \{\mathrm{y} : (\mathrm{x}, \mathrm{y}) \in \mathrm{T}\}$ is the set of all technologically feasible output vectors given inputs x, with outer boundary given by the output isoquant $\mathrm{I}(x) = \{\mathrm{y} \in \mathrm{P}(\mathrm{x}), \lambda\mathrm{y} \notin \mathrm{P}(\mathrm{x}) \forall \lambda > 1\}$. The output distance function is then defined on P(x) as

$$D_o(x, y) = \min\{\lambda : (y/\lambda) \in P(x)\}$$

[2.64]

It is conventional to define the output oriented Malmquist productivity index as the geometric mean of the two Malmquist productivity indexes defined in period t and t+1, written as

$$\begin{aligned} M_o\left(x^t, y^t, x^{t+1}, y^{t+1}\right) &= \left\{M_o^t\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times M_o^{t+1}\left(x^t, y^t, x^{t+1}, y^{t+1}\right)\right\}^{1/2} \\ &= \left\{\frac{D_o^t\left(x^{t+1}, y^{t+1}\right)}{D_o^t\left(x^t, y^t\right)} \times \frac{D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_o^{t+1}\left(x^t, y^t\right)}\right\}^{1/2} \end{aligned}$$

[2.65]

As both indexes $M_o^t\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ and $M_o^{t+1}\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ compare $\left(x^{t+1}, y^{t+1}\right)$ and

(x^t, y^t) based on production technologies between period t and t+1, the choice of production technology is totally arbitrary and these two indexes are not necessarily equal to each other.

The main debate about the Malmquist productivity index shown in [2.65] lies in the fact that although Caves *et al.* (1982) defined their index on a technology that allows for varying returns to scale. It actually ignored the contribution of scale economies to productivity growth, as argued by Grifell-Tatjé and Lovell (1995). Färe and Grosskopf (1996) also proved that for the single output and single input case, the Caves *et al.* index provides an accurate measure of productivity change if, and only if, the index is defined on a technology exhibiting constant returns to scale. As suggested by Fried *et al.* (2008), it is now the common sense to define the Caves *et al.* index on a benchmark technology satisfying constant returns to scale as to distinguish from the Malmquist productivity index defined on a best practice technology allowing for variable returns to scale, which enables the Malmquist productivity index to account for the effect of scale economies on the productivity change, as a departure of the best practice technology from the benchmark technology. This convention is also very important for researchers interested not only in the overall productivity measurement but also in the potential sources of the productivity growth to construct the correct and meaningful decomposition of the productivity change.

To redefine Caves *et al.* (1982) Malmquist productivity index on a benchmark technology satisfying constant returns to scale, assuming that the firm use inputs $x \in \Re_+^M$ to produce outputs $y \in \Re_+^J$ with the benchmark technology set that $\mathrm{T_b} = \{(\mathrm{y}, \mathrm{x}) : \mathrm{x} \text{ can produce y}\}$ under conditions of global constant returns to scale. The output distance function is then redefined on P(x) as

$$D_{ob}(x, y) = \min\{\lambda : (y/\lambda) \in P(x)\} \qquad [2.66]$$

and the output-oriented Malmquist productivity index with respect to benchmark technology is redefined as

$$
\begin{aligned}
M_{ob}\left(x^t, y^t, x^{t+1}, y^{t+1}\right) &= \left\{M_{ob}^t\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times M_{ob}^{t+1}\left(x^t, y^t, x^{t+1}, y^{t+1}\right)\right\}^{1/2} \\
&= \left\{\frac{D_{ob}^t\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^t\left(x^t, y^t\right)} \times \frac{D_{ob}^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^{t+1}\left(x^t, y^t\right)}\right\}^{1/2}
\end{aligned}
$$

[2.67]

where the notation ‘ $_{ob}$ ’ ,‘ $^{t}_{ob}$ ’ and ‘ $^{t+1}_{ob}$ ’ indicate the distance functions comprising the Malmquist productivity index defined on the period t and t+1 benchmark technology, respectively.

As mentioned earlier, economic researchers or industry regulators are not only interested in measuring the overall productivity growth, but also finding the possible reasons behind such productivity growth, for instance, whether the productivity growth is driven by the improvement of firm’s overall performance and executives’ ability, reflected as the improvement of technical efficiency; or by the utilization of advanced technology in the production process, known as technical progress; or by the reduction in unit cost during business expansion, indicating the benefits from scale effect change; or adjustment from using the correct input and output mix as allocative efficiency gains. Thus the motivation to find sources and decomposition of productivity change is deep-seated in economics.

The first decomposition of Malmquist productivity index, written in [2.68] is addressed by Färe *et al.* (1994a) as

$$
\begin{aligned}
M_{ob}\left(x^t, y^t, x^{t+1}, y^{t+1}\right) &= \left\{\frac{D_{ob}^t\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^t\left(x^t, y^t\right)} \times \frac{D_{ob}^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^{t+1}\left(x^t, y^t\right)}\right\}^{1/2} \\
&= \left[\frac{D_{ob}^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^t\left(x^t, y^t\right)}\right] \times \left\{\frac{D_{ob}^t\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^{t+1}\left(x^{t+1}, y^{t+1}\right)} \times \frac{D_{ob}^t\left(x^t, y^t\right)}{D_{ob}^{t+1}\left(x^t, y^t\right)}\right\}^{1/2} \\
&= TE\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)
\end{aligned}
$$

[2.68]

where $TE\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ measures the contribution of technical efficiency change to productivity change. $TE\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ greater, equal to or less than one indicates that technical efficiency improves, remains unchanged or deteriorates between periods t and t+1. $T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ measures the contribution of technical change to productivity change as the geometric mean of two terms, one comparing period t benchmark technology to period t+1 benchmark technology from the perspective of period t data, and the other comparing the two benchmark technologies from the perspective of period t+1 data. $T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ greater, equal to, or less than one suggests technical progress, stagnation, or regress occurred between periods t and t+1.

However, this initial decomposition has a problem since productivity change is measured relative to the benchmark technology that is constrained to constant returns to scale, and so is its components technical efficiency change and technical change. It ignores the contribution of scale economies to productivity change that can only be recognized under the best practice technology allowing for various returns to scale. Therefore, productivity change and its decomposition need to be redefined under the best practice technology.

To do so, Färe *et al.* (1994b) redefined technical efficiency change component in their initial decomposition shown in [2.68] to obtain

$$
\begin{aligned}
TE\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right) &= \left[\frac{D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_o^t\left(x^t, y^t\right)}\right] \times \left\{\frac{D_{ob}^{t+1}\left(x^{t+1}, y^{t+1}\right) / D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^t\left(x^t, y^t\right) / D_o^t\left(x^t, y^t\right)}\right\}^{1/2} \\
&= TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times \left[\frac{SE^{t+1}\left(x^{t+1}, y^{t+1}\right)}{SE^t\left(x^t, y^t\right)}\right] \\
&= TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)
\end{aligned}
$$

[2.69]

where $TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ measures technical efficiency change on the best practice technologies and $SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ measures the change in scale efficiency from period t to t+1. Inserting [2.69] back into [2.68] gives the Färe *et al.* (1994b) decomposition of the Malmquist productivity index

$$
\begin{aligned}
M_{ob}\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = {} & TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \\
& \times T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)
\end{aligned}
$$

[2.70]

Ray and Desli (1997) were the first of many to criticize this decomposition. Since the magnitude of a shift in the benchmark technology has little to do with the magnitude of a shift in the best practice technology, although the first part of Färe *et al.* (1994b) decomposition, $TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$, is measured relative to the best practice technology, the technical change component $T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$, defined relative to the benchmark technology, can overstate or understate the magnitude of technical change on the best practice technology. Hence there must be something else in the component of technical change. Also, the scale efficiency change is measured on quantity vectors and technology between period t and t+1. Hence it must combine the effects of scale economies and technical change. As suggested in Lovell (2003), a component of productivity change is missing from [2.70], and its contribution is embedded in $SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ and $T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$.

Ray and Desli (1997) attempted to remedy Färe *et al.* (1994b) decomposition by isolating $T\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$, the shift in the best practice technology through the effort of merging $SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ with $T\Delta_b\left(x^t, y^t, x^{t+1}, y^{t+1}\right)$ in [2.70] to obtain the alternative decomposition

$$\begin{aligned} M_{ob}\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \times SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \\ \times T\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \end{aligned}$$

[2.71]

Where $$TE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = \frac{D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_o^t\left(x^t, y^t\right)} \qquad [2.72]$$

$$T\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) = \left[\frac{D_o^t\left(x^{t+1}, y^{t+1}\right)}{D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)} \times \frac{D_o^t\left(x^t, y^t\right)}{D_o^{t+1}\left(x^t, y^t\right)}\right]^{1/2} \qquad [2.73]$$

and the scale change factor

$$\begin{aligned} & SE\Delta\left(x^t, y^t, x^{t+1}, y^{t+1}\right) \\ & = \left\{\left[\frac{D_{ob}^t\left(x^{t+1}, y^{t+1}\right) / D_o^t\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^t\left(x^t, y^t\right) / D_o^t\left(x^t, y^t\right)}\right] \times \left[\frac{D_{ob}^{t+1}\left(x^{t+1}, y^{t+1}\right) / D_o^{t+1}\left(x^{t+1}, y^{t+1}\right)}{D_{ob}^{t+1}\left(x^t, y^t\right) / D_o^{t+1}\left(x^t, y^t\right)}\right]\right\}^{1/2} \\ & = \left[\frac{SE^t\left(x^{t+1}, y^{t+1}\right)}{SE^t\left(x^t, y^t\right)} \times \frac{SE^{t+1}\left(x^{t+1}, y^{t+1}\right)}{SE^{t+1}\left(x^t, y^t\right)}\right]^{1/2} \end{aligned}$$

[2.74]

The Ray and Desli (1997) efficiency change term is the same as that of Färe *et al.*

(1994b), but their technical change term differs from that of Färe *et al.* (1994b) as it is defined on the best practice technology. Their scale change factor is the geometric mean of a pair of scale efficiency ratios, measured on period t technology and the other measured on period t+1 technology. Hence the scale change factor refers to the quantity vectors but not the technology.

Although Färe *et al.* (1997) criticize that Ray and Desli (1997) scale change factor does not actually measure scale efficiency change since each component uses only a single period technology, however, Lovell (2003) claims that Färe *et al.* (1997) criticism is "*mathematically correct but economically misguided*" and the term "scale efficiency change" should not be considered as a source of productivity change, which he believes "*has misled researchers for years*". Returning from original Caves *et al.* (1982) formulation of geometric mean of Malmquist productivity index, Lovell derives the similar decomposition as Ray and Desli (1997) along with two additional economically meaningful decompositions (see Lovell, 2003).

To summarize, the Malmquist productivity index has several nice theoretical features. Because it is based on distance functions, it inherits several desirable properties from them and can handle multiple outputs and multiple inputs. It can be easily measured in an output orientation as output oriented Malmquist productivity index but also as an input oriented Malmquist productivity index based on input distance.

The Malmquist productivity index also has a very nice practical feature. Since it is based on distance functions, it only requires information on output and input quantities, but not the output and input prices. This makes it an extremely applicable tool for productivity measurement in situations that prices are distorted or missing.

Furthermore, the Malmquist productivity index can be decomposed into economically meaningful sources of productivity change, as discussed extensively above. This also meets the economic incentive not only to measure the overall productivity growth but also to identify the sources and origins. By using the mathematical programme DEA, it is not difficult to construct the best practice production benchmark and calculate the productivity change and its sources. However, because it is based on DEA, it shares its shortcomings such as ignorance of random noise and incapability of provide statistical significance test. Once again, since it is based on DEA, it cannot avoid the debate of comparisons between DEA and SFA and the criticism from researchers who favour SFA. The pros and cons of DEA and SFA have been discussed in section 2.2. No consensus on this debate leaves a large number of researchers unpersuaded by either side of the debate but also justifies the use of parametric frontier approaches to measure the productivity change.

2.3.2.2.2 Parametric frontier approaches

Compared to the large volume of productivity studies using non-parametric DEA approach to measure the productivity change using the Malmquist productivity index, the literature of parametric approaches is smaller. As argued in Grosskopf (1993), the first effort to calculate technical change using a parametric frontier model was that of Førsund and Hjalmarsson (1979). They applied a linear programming approach developed by Aigner and Chu (1968) to calculate a production frontier. They generalized the production frontier to the homothetic case, and included time trends on purpose to identify frontier technical change. They solved a linear programming problem of the following general form:

$$\min \sum_{t=1}^{T} \sum_{k=1}^{K} \left[\ln A + \gamma_t t + \gamma_{lt} t \ln L_k(t) + \gamma_{kt} t \ln K_k(t) - \gamma_{yt} t \ln y_k(t) \right] \qquad [2.75]$$

subject to

$$\ln A + \gamma_t t + \gamma_{lt} t \ln L_k(t) + \gamma_{kt} t \ln K_k(t) - \gamma_{yt} t \ln y_k(t) \geq 0$$
$$t = 1,\dots,T \quad k = 1,\dots,K$$

subject also to nonnegativity constraints and homogeneity constraints. L and K represents labour and capital, respectively. The objective is to minimize the sum of deviations of frontier from observed output. With the restrictions of nonnegativity constraints, it indeed treats all the deviations from the frontier as inefficiency, that is, the model constructs a deterministic, parametric frontier with no allowance for other types of error.

Nishimizu and Page (1982) used an approach similar to Førsund and Hjalmarsson (1979) but with two exceptions. First, they specified the production technology as translog. Secondly, although Førsund and Hjalmarsson focused on isolating frontier technical change, they focus on measuring productivity growth, identified as the sum of frontier technical change and change in efficiency, much like that developed for the Malmquist productivity index.

Perelman and Pestieau (1988) extended Nishimizu and Page's work by estimating a deterministic production function in the case of scalar output. They included a time trend in the specification and corrected ordinary least squares was used to estimate the frontier. Following Nishimizu and Page, they defined the same productivity index as technical change and changes in efficiency. But unlike Nishimizu and Page, they estimated technical change as the derivatives of the production frontier with respect to time evaluated at year t. Thus their index and its decomposition is the parametric, continuous counterpart of the non-parametric, discrete time measure of Malmquist productivity index and its decomposition.

Since then, Perelman and Pestieau's (1988) parametric deterministic approach has been generalized to a stochastic frontier and applied into industries and financial service sectors. Fecher and Perelman (1992) and Fecher and Pestieau (1993) used the production function with a scalar output to calculate the technical change and productivity growth. However, the production process associated with multiple outputs are more likely in the real world and this reality requires other functions than production function to represent the technology and measure productivity.

One extension fulfilling this desire was proposed by Bauer (1990), in which the author calculated TFP change using a stochastic cost frontier. His expression decomposes TFP growth into terms relate to returns to scale, technical and allocative efficiency change, technical progress and input price effect. Detailed discussion and derivation of Bauer's approach will be provided in Chapter 5.

In another extension Orea (2002) provided a parametric decomposition of the generalized Malmquist productivity index using a distance function approach. In his approach, the firm's technology is represented by a translog output-oriented distance function, written as

$$
\begin{aligned}
\ln D_o(x,y,t) = {} & \alpha_0 + \sum_{m=1}^{M}\alpha_m \ln x_m + \sum_{j=1}^{J}\beta_j \ln y_j + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}\alpha_{mn}\ln x_m \ln x_n \\
& + \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J}\beta_{jk}\ln y_j \ln y_k + \sum_{m=1}^{M}\sum_{j=1}^{J}\gamma_{mj}\ln x_m \ln y_j \\
& + \theta_1 t + \frac{1}{2}\theta_2 t^2 + \sum_{m=1}^{M}\xi_m t \ln x_m + \sum_{j=1}^{J}\zeta_j t \ln y_j
\end{aligned}
$$

[2.76]

Since this distance function [2.76] can be regarded as a quadratic function in variables x, y and t, by using Diewert's (1976) Quadratic Identity Lemma, changes of distance function [2.76] from period t to t+1 can be written as

$$
\begin{aligned}
& \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right) - \ln D_o\left(x^t, y^t, t\right) \\
& \quad = \frac{1}{2}\sum_{j=1}^{J}\left[\frac{\partial \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right)}{\partial \ln y_j} + \frac{\partial \ln D_o\left(x^t, y^t, t\right)}{\partial \ln y_j}\right]\cdot\left(\ln y_j^{t+1} - \ln y_j^t\right) \\
& \quad + \frac{1}{2}\sum_{m=1}^{M}\left[\frac{\partial \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right)}{\partial \ln x_m} + \frac{\partial \ln D_o\left(x^t, y^t, t\right)}{\partial \ln x_m}\right]\cdot\left(\ln x_m^{t+1} - \ln x_m^t\right) \\
& \quad + \frac{1}{2}\left[\frac{\partial \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right)}{\partial t} + \frac{\partial \ln D_o\left(x^t, y^t, t\right)}{\partial t}\right]
\end{aligned}
$$

[2.77]

A logarithmic Malmquist productivity index $\ln M_o(x,y,t)$ can be defined as

$$
\begin{aligned}
\ln M_o(x,y,t) = {} & \frac{1}{2}\sum_{j=1}^{J}\left[\frac{\partial \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right)}{\partial \ln y_j} + \frac{\partial \ln D_o\left(x^t, y^t, t\right)}{\partial \ln y_j}\right]\cdot\left(\ln y_j^{t+1} - \ln y_j^t\right) \\
& - \frac{1}{2}\sum_{m=1}^{M}\left[\frac{-\partial \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right)}{\partial \ln x_m} + \frac{-\partial \ln D_o\left(x^t, y^t, t\right)}{\partial \ln x_m}\right]\cdot\left(\ln x_m^{t+1} - \ln x_m^t\right)
\end{aligned}
$$

[2.78]

which is the difference between weighted average rates of outputs and inputs growth with the weights of outputs and inputs distance elasticities derived from partial differentiation of distance function with respect to outputs and inputs, respectively.

Combining [2.77] and [2.78], $\ln M_o(x,y,t)$ is decomposed as

$$\ln M_o(x,y,t) = \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right) - \ln D_o\left(x^t, y^t, t\right)$$
$$- \frac{1}{2}\left[\frac{\partial \ln D_o\left(x^{t+1}, y^{t+1}, t+1\right)}{\partial t} + \frac{\partial \ln D_o\left(x^t, y^t, t\right)}{\partial t}\right]$$

[2.79]

This expression [2.79] decomposes $\ln M_o(x,y,t)$ into two meaningful parts, technical efficiency change and technical change. This decomposition can be viewed as the parametric counterpart of the traditional output oriented DEA-type Malmquist productivity index as shown in [2.68] when the output distance function is specified as translog. However, as a general consensus, the TFP index should satisfy four desirable properties: identity, monotonicity, separability and proportionality. The first identity property implies that if inputs and outputs do not change from period t to t+1, the TFP change equal to one. The monotonicity property requires that a TFP index is constructed with higher output and lower input usage indicating the improvement in TFP. Separability implies that a TFP index is interpreted in the same way as in the single-output single-input case, for example, in the multiple-outputs multiple-inputs case, the aggregated output growth rate only relies on the output data and the aggregated input growth rate only depends on the input data. Therefore, if the technology is separable in outputs and inputs, the TFP index has the desirable property. The proportionality property suggests that the weights in the output and input growth indices should add to one. $\ln M_o(x,y,t)$ fulfills the first three but fails the last one since the inputs weights in [2.78] do not sum to one. Therefore, in the case of various returns to scale, $\ln M_o(x,y,t)$ may not be a TFP index since it ignores the potential contribution of scale economies to productivity change.

To extend the decomposition of $\ln M_o(x,y,t)$ to allow for the effect of various returns to scale, Orea drew on the ideas suggested by Denny *et al.* (1981) and developed a generalized Malmquist productivity index that can incorporate the scale effect. Using output and input distance elasticity shares rather than distance elasticities, a generalized output-oriented Malmquist productivity index can be defined as

$$\ln G_o(x,y,t)=\frac{1}{2}\sum_{j=1}^{J}\left[\varepsilon_j\left(x^t,y^t,t+1\right)+\varepsilon_j\left(x^t,y^t,t\right)\right]\cdot\left(\ln y_j^{t+1}-\ln y_j^t\right)$$
$$-\frac{1}{2}\sum_{m=1}^{M}\left[e_m\left(x^t,y^t,t+1\right)+e_m\left(x^t,y^t,t\right)\right]\cdot\left(\ln x_m^{t+1}-\ln x_m^t\right)$$

[2.80]

where $\varepsilon_j\left(x^t,y^t,t\right)=\dfrac{\partial \ln D_o\left(x^t,y^t,t\right)}{\partial \ln y_j}$

$$e_m\left(x^t,y^t,t\right)=\frac{\partial \ln D_o\left(x^t,y^t,t\right)/\partial \ln x_m}{\sum_{m=1}^{M}\partial \ln D_o\left(x^t,y^t,t\right)/\partial \ln x_m}$$

Since input and output weights are now sum to one, $\ln G_o(x,y,t)$ satisfies the proportionality and can be viewed as a TFP index in the case of various returns to scale. Using [2.77] and [2.79], the generalized Malmquist productivity index in [2.80] can be decomposed into $\ln M_o(x,y,t)$ and a term of returns to scale. That is:

$$\ln G_o(x,y,t)=\ln M_o(x,y,t)+\frac{1}{2}\sum_{m=1}^{M}\left[\left(-\sum_{m=1}^{M}\frac{\partial \ln D_o\left(x^{t+1},y^{t+1},t+1\right)}{\partial \ln x_m}-1\right)\cdot e_m\left(x^t,y^t,t+1\right)\right.$$
$$\left.+\left(-\sum_{m=1}^{M}\frac{\partial \ln D_o\left(x^t,y^t,t\right)}{\partial \ln x_m}-1\right)\cdot e_m\left(x^t,y^t,t\right)\right]\cdot\left(\ln x_m^{t+1}-\ln x_m^t\right)$$

[2.81]

All that required to calculate the $\ln G_o(x,y,t)$ and its three sources is to estimate the translog output-oriented distance function [2.77], imposing homogeneity in outputs and making an assumption about the error structure. Then parameter estimates can be used to calculate technical efficiency change, technical change, the output and input distance elasticities and the output and inputs distance elasticity shares required to calculate the scale effect change. The TFP index is the sum of these three components.

To compare Bauer (1990) approach and Orea (2002) approach, both of them use a parametric stochastic frontier approaches that allow for random noise to measure the TFP change and provide a method to decompose the TFP index that allow researchers to identify the potential sources of the productivity change. Also, both approaches are built on the production technology that allows for multiple outputs.

However, despite the common factors, these two approaches differ from each other in the perspectives of both theoretical framework and empirical application. Regarding the theoretical built-up, Bauer (1990) adopted a stochastic cost function while Orea (2002) utilized a stochastic output-oriented distance function. Due to availability of input and output prices information, Bauer managed to decompose the TFP into five components: scale effect change, cost efficiency change that could further decomposed into technical efficiency change and allocative efficiency change, technical change, input price effect term and output price effect terms arising from the nonmarginal output pricing effect. While, without requirement of information on input and output price, Orea (2002) provides a generalized TFP index and decomposes it into technical efficiency change, technical change and change in returns to scale. What is missing in Orea's (2002) decomposition is the allocative efficiency change effect, which could come from both input and output mix. From the perspective of empirical practice, Bauer's (1990) approach is constructed on a continuous time, known as total differential approach, while Orea's (2002) approach is based on discrete time, known as index number approach. As suggested in Coelli *et al.* (2003), the main difference between these two approaches lies in that the former choose just one data point for derivative evaluation while the latter approach can evaluate derivatives at both data points. In the other word, the total differential approach allows one to calculate the productivity change for the whole sample period, but the index number approach can allow one to evaluate the productivity change for the separate sample points over the sample period. This difference makes Orea's (2002) approach applicable in the empirical literature that usually involves comparison of productivity change in different sample points.

Therefore, this thesis is motivated to develop an index number counterpart of Bauer's (1990) approach using the stochastic cost frontier approach. Moreover, this index number cost frontier approach also has another desirable feature to allow the researchers to incorporate the effect of allocative efficiency change arisen from wrong input and output usage. Hence, this thesis is among the first to measure the TFP change and its decomposition using a parametric stochastic cost frontier approach based on discrete time with the application to ten major Asian banking industries.

2.3.3 Summary

The purpose of this small subsection is to provide a brief overview of approaches to measure productivity and identify its sources. This brief overview is implemented through two dimensions: non-frontier and frontier approaches, which have been further subdivided into non-parametric and parametric models. The non-parametric frontier approach involves programming techniques, such as DEA, while the parametric frontier approaches include both the stochastic and deterministic models.

Comparing the frontier approaches to the non-frontier approaches, the latter has the advantage of simplicity and computational ease. They do not need to estimate any parameters or run any linear programming technique. However, they are vulnerable to bias since these approaches are based on the assumption of no technical and allocative inefficiency. Within the non-frontier approaches, the non-parametric approaches have the advantage relative to their parametric counterparts of avoiding any specification error, albeit at the expense of ignoring the random noise that leads to measurement error and sampling error.

Comparing the parametric frontier approaches to the non-parametric frontier approaches, the latter has the advantage of minimal specification error, but does not allow for measurement error or random shocks. The specification error can be raised from either the assumption of functional form or from the specification of the distribution of the error term. However, the problem of specification error can be minimized by using flexible functional form and test of theoretical properties associated with those functional forms. Within the parametric frontier approaches, the stochastic frontier approach has the advantage of allow for the random shocks and measurement error.

Chapter 3 Purpose of the Empirical Work and the Data[4]

3.1 Purpose of the empirical work

Motivation of this thesis lies in the fact that banking efficiency and productivity studies have been dominated by those studies that evaluate country-specific efficiency levels and productivity changes, conduct cross-country comparisons and assess effects of deregulation, privatization, and consolidation in developed countries such as the US and European Union. However only limited Asian studies exist, which mainly focus on individual banking industry with only one exception of cross-country study in four South Asian countries (Perera *et al.* 2007). Therefore, I am motivated to fill this literature gap with an attempt to measure and compare the efficiency and productivity change of banking industries in major Asian economies.

---

[4] Data from this chapter has been used in the following two works published by Taylor & Francis: Shen, Z., Liao, H. and Weyman-Jones, T. 2009, "Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions", *Journal of Chinese Economic and Business Studies*, vol. 7, no. 2, pp 199-218.
Shen, Z., Liao, H. and Weyman-Jones, T. 2009, "Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions" in *China's Three Decades of Economic Reforms*, eds. X.H. Liu and W. Zhang, Routledge, Taylor & Francis Group, UK.

In banking efficiency literature, one can barely find any frontier efficiency studies using production function (as seen in my survey, Berger *et al.*, 1993 and Berger and Humphrey, 1997). It is mainly because production function can only be employed in the case of one output, while in the banking industry, banks usually provide different kinds of services that generate multiple outputs, such as loans, other earning assets, and off-balance sheet income. Even when only loans are considered as output, it can still have several outputs in regard to different characteristics of loans, such as, in terms of loan purpose, consumer loans, real estate loans, interbank loans, and also as, in terms of loan maturity, short term loans and long term loans. When these data are available, researchers will no doubt find production frontier analysis inapplicable and seek to find the alternative ways to fulfil their research interests. Therefore, instead of measuring technical efficiency that is related to production function, scholars measure X-efficiency, such as cost efficiency and profit efficiency.

Measurement of cost efficiency, as discussed in the last chapter, requires data on level of outputs and input prices, whereas standard measurement of profit efficiency needs data on output and input prices. However, in most banking efficiency and productivity studies, information on output prices is hardly tenable since it is difficult to find some of the banks' output prices in balance sheet and income statements, for instance, prices of investments in securities or off-balance sheet non-interest income. This difficulty justifies the emergence of a large number of cost efficiency studies and non standard measurement of profit efficiency that employs the same independent variables, outputs and input prices, as in measuring cost efficiency but replacing the dependent variable, a measure of cost, with a measure of profit. However, although non standard measure of profit efficiency could provide useful information in some occasions as discussed in section 2.2.2.4, its satisfaction of theoretical properties associated with profit function has never been checked. Therefore, it's better to measure cost efficiency than to give any misleading profit efficiency estimates obtained from misspecifications of the true standard profit function.

It has been widely addressed in the efficiency literature that cross-country comparisons should take into account the impact the cross-country environmental variables. However, two questions can be addressed on this issue. First, most of these studies reach this general consensus by employing cross-sectional stochastic frontier analysis.

Whether this consensus still stands in the panel data stochastic frontier approaches requires further evidence. Second, if so, how these environmental variables influence the specified frontier model is also worth discussion. The existing literature shares the same assumption that cross-country environmental variables will have an impact on model structure that they will influence the shape of the stochastic frontier. However, as suggested in Coelli *et al.* (1999), exogenous variables such as these cross-country environmental variables may not only influence the shape of the frontier but also influence the mean of inefficiency. Efficiency level obtained from the former model is termed as net efficiency that is net of environmental influences, while the one estimated from the latter model, is named as gross efficiency. To extend Coelli *et al.*'s (1999) argument, environmental influences could also enter the variance of inefficiency or the variance of random errors as a measure of heteoscedasticity.

This thesis will address the above research questions in the next three empirical chapters. Chapter 4 will measure and compare cost efficiency in ten Asian banking industries using advanced panel data stochastic frontier approaches with incorporation of cross-country environmental variables as a control of the shape of cost frontier. The main purpose of chapter 4 is to assess the appropriateness of different panel data models in efficiency studies, and to determine whether cross-country environmental variables are truly an important factor to explain part of inefficiency obtained from those models that ignore these cross-country heterogeneities.

Based on the efficiency and parameter estimates of the best model from chapter 4, the TFP change in these ten Asian banking industries is measured in chapter 5 by adopting a new parametric index number approach that decomposes the TFP change into five components including the impact of output and input mix allocative efficiency change. This new cost based parametric index number approach would be an important addition to productivity literature that has been dominated by the DEA type Malmquist productivity index approach because it provides some new sources to explain the total factor productivity change.

Chapter 6 examines the exogenous influences of environmental variables on the composed error terms, by either including them in the mean of inefficiency as determinants, or incorporating them in the variance of inefficiency term as a measure to control heteroscedasticity, or modelling them in the variance of random error term as another measure of heteroscedasticity. Several hypotheses are tested in terms of choosing the best model and addressing the correct effects of exogenous influences with further estimation of country specific cost efficiency and calculation of the TFP change.

3.2 Data

The main task of this research is to measure and compare efficiency and productivity change in the Asian banking industries using panel data stochastic cost frontier approaches considering the impacts of exogenous environmental variables. But first and foremost, sample data should be constructed.

Living commercial banks from ten Asian countries and regions (listed as China, Hong Kong SDR, India, Indonesia, Korea, Malaysia, Philippines, Singapore, Taiwan Province of China and Thailand[5]) are included in this study. Based on the generally accepted accounting principles (GAAP), annual unconsolidated bank data from 1998 to 2005 are collected from the Bankscope database that provides detailed financial information for 11,000 banks over the world. Currently only those commercial banks that focus on retail banking business and have a positive value of deposit loans margin are considered. As distinct from wholesale commercial banks, retail commercial banks work as financial intermediaries that collect deposits from customers and loan out to investors. The interest margin between loans and deposits therefore provides most of the banks' income. To maximize shareholders' profits, some bank managers may expand their loans beyond the limits of deposits, behaving like investment banks and therefore using borrowing from wholesale markets to replace deposits. Given a sound economic background and booming financial markets, these banks would benefit from their loan expansion and enjoy superior efficiency of banking performance. However, in the long run, they could face high default risks, causing huge losses and even bankruptcy, as has been apparent in the credit difficulties of 2008.

For those missing data not reported in Bankscope, they are collected from alternative

[5] The sampling countries and regions are neighboring economies that are more or less influenced by the 1997 Asian financial crisis. To focus on them could help us to have in-depth understanding of the impact of financial crisis on banks' performance.

data sources, for instances, the annual report of individual bank, the statistical yearbook of individual country, the statistical department of individual country and the labour department of individual country. To maintain the consistency with the data from Bankscope, all data from alternative sources are carefully checked.

Since all banks' data are in own-currency nominal values, to allow the cross-country comparison, they are covented to US dollars using a purchase power parity exchange rate (PPP) that contains an adjustment for inflation (see Table 3.1). Therefore, it ends up with an unbalanced panel data set consisting of 280 banks (48 Chinese banks, 22 banks in Hong Kong SDR, 56 Indian banks, 35 Indonesian banks, 21 Malaysian banks, 23 Philippine banks, 3 Singaporean banks, 16 Korean banks, 41 banks in Taiwan province of China and 15 banks in Thailand) with a total number of 1890 observations.

Table 3.1: PPP exchange rate for the Asian countries (National currency per current international dollar)

Country	1998	1999	2000	2001	2002	2003	2004	2005
China	2.011	1.966	1.964	1.957	1.935	1.944	2.021	2.047
Hong Kong	8.666	8.051	7.436	7.129	6.763	6.202	5.811	5.63
India	8.034	8.256	8.386	8.488	8.606	8.778	8.894	9.064
Indonesia	1,827.49	2,056.52	2,184.84	2,438.68	2,538.15	2,621.65	2,766.69	3,084.97
Korea	769.438	757.709	747.013	755.32	763.305	767.765	766.653	744.381
Malaysia	1.622	1.6	1.641	1.557	1.587	1.609	1.662	1.687
Philippines	9.739	10.371	10.794	11.214	11.518	11.704	12.065	12.442
Singapore	1.712	1.598	1.623	1.554	1.512	1.467	1.481	1.446
Taiwan	21.392	20.81	20.037	19.663	19.163	18.362	17.57	16.957
Thailand	13.354	12.631	12.529	12.488	12.374	12.29	12.344	12.527

Source: International Monetary Fund, World Economic Outlook Database, April 2007

Table 3.2 summarizes total assets shares of sampled banks in the whole banking system. The majority of total assets shares are over 80% during the sample period despite the fact that for countries like India, Korea, Malaysia, Philippines and Thailand, this figure is relative small at the beginning of the sample period, especially for 1998. This gap can be explained from two aspects. First of all, it may be caused by lack of sufficient bank observations. For example, in 1998, only 2 (out of 56) Indian banks are observed. The same reason stands for Thailand as 3 (out of 15) banks are observed. Secondly, the relative large gap at the beginning of the sample period may also partly be explained by the impact of the 1997 Asian financial crisis. In the post crisis period, large number of banks are forced out of the market or required to merge with large and healthy banks. For example, at the end of 1997, there are 223 Indonesian banks while in June 2002, only 145 banks left. Moreover, by the end of 2005, only 19 banks exist in Korea compared with the figure of 33 banks at the end of 1997. The nature of this sample that only takes account of living commercial banks may cause this considerable gap in total asset shares. But nevertheless, the fact that total asset shares are around 75-90% for the majority of the sample period suggests that this sample covers the majority in the banking system and empirical findings will present this big picture.

Table 3.2: Total assets shares for sampling banks in the banking system (1998-2005)

	China	Hong Kong	India	Indon esia	Korea	Malay sia	Philip pines	Singa pore	Taiwa n	Thaila nd
1998	86.0	87.9	7.9	81.8	59.3	34.8	66.8	65.0	80.3	32.7
1999	86.0	71.3	79.5	90.8	80.4	64.7	78.3	62.6	80.7	83.1
2000	86.6	74.5	78.5	89.8	81.6	71.0	88.0	61.8	79.6	82.7
2001	93.5	78.1	77.9	92.9	91.9	71.3	91.1	82.2	80.4	81.2
2002	89.6	81.5	83.7	94.0	95.7	69.9	90.5	79.1	81.2	83.0
2003	89.8	83.5	83.6	90.2	96.7	73.4	88.9	90.1	82.6	83.3
2004	88.0	84.7	87.8	85.3	98.0	76.0	91.9	90.9	80.3	86.9
2005	87.3	85.8	85.8	83.2	96.3	80.3	90.8	90.0	80.3	86.6
No. of banks	48	22	56	35	21	23	3	16	41	15

Source: Bankscope Database, 2007 and authors' own calculation.

Table 3.3: Total assets share of Top 10 commercial banks (1998-2005)

	China	Hong Kong	India	Indon -esia	Korea	Malay - sia	Philipp - ines	Singap- ore	Taiwan	Thailand
1998	83.3	79.0	-	62.6	53.0	23.6	51.6	65.0	52.4	26.1
1999	82.3	64.2	47.6	78.9	76.8	52.7	61.6	62.6	52.2	79.5
2000	81.2	68.7	46.4	81.4	79.0	57.1	71.6	61.8	52.0	79.3
2001	86.4	71.9	46.4	83.2	88.7	56.0	79.3	82.2	52.4	77.9
2002	81.4	74.7	51.7	81.9	91.2	58.8	74.5	79.1	52.1	79.8
2003	80.7	76.8	50.9	76.1	92.1	61.5	72.8	90.1	53.2	79.9
2004	80.1	78.2	51.4	69.7	93.1	64.8	74.9	90.9	50.7	83.1
2005	79.9	79.1	50.3	66.7	91.1	64.6	76.0	90.0	50.3	82.1

Notes: 1. No ratio for Indian banks in 1998 due to lack of data; 2. Since only three Singaporean banks are included in the sample, total asset shares for Top 10 Singaporean banks are actually measured as Top 3 banks.

Source: Bankscope Database, 2007 and authors' own calculation.

It is also of my interest to study the structure of the banking system, especially the banking concentration and market power. As seen in Table 3.3, Top 10 commercial banks in individual country and region, ranked by their total assets, possess the majority of total assets in the banking system. Top 10 commercial banks in China, Hong Kong SDR, Korea, Philippines, Singapore and Thailand has a total market share of over 70%, indicating the dominance of large commercial banks in the banking system. However, Top 10 banks from India, Malaysia and Taiwan Province of China have a lower market share of about 50 to 60 percent, indicating the relative competitive banking market in these countries. These descriptive data are consistent with the data of banking concentration reported in IMF database on Financial Development and Structure (see Table 3.6), where the banking concentration ratio for India, Malaysia and Taiwan Province of China is 36%, 44% and 30% respectively.

Consequently, although the sample data set does not include all the commercial banks of individual country, it does cover the majority of the banking system and present the top tier. And the measurement of banking performance, cost efficiency and the productivity change from using this sample will reflect the magnitude of the banking activity of those Asian countries and give us a major image of how well the whole banking system is running.

3.2.1 Dependent and Independent Variables

Table 3.4 and Table 3.5 summarize the dependent and independent variables in this study. In next section, we will discuss how each variable is defined and the way it is measured and processed.

The definition of total costs is termed as the sum of the interest expense, the personnel expenses and other operating expenses.

Regarding the definition of outputs and input specification, the intermediation approach is adopted. As discussed in section 2.1.2.2.1, there is no consensus about which approach should be adopted to define the inputs and outputs. The centre of the debate is the role of deposits that whether it should be regarded as inputs or outputs. Production approach considers banks as a production unit to provide financial services to the customers. Therefore the deposits' account provided to the customers should be included as one of the outputs of banks. However, the counterpart intermediation approach views banks as financial intermediaries collecting funds from investors and loaning them out to earn the interest margin. Then deposits are considered as inputs. Unfortunately, neither of the two approaches fully captures the dual role of the financial institutions. Therefore, as argued by Berger and Humphrey (1997), deposits should be considered as both inputs and outputs and a dual approach should be used. However, although dual approach is well explained in theory and applied in empirical studies, there are no statistical tests to check whether certain properties stated in the theory of production function, or its dual cost function and profit function, are satisfied when dual approach is used. In my knowledge so far, this issue has not been discussed and dual approach is used as granted. Therefore, there is room for me to provide statistic test to compare and discuss these distinct output and input specifications.

In this data demonstration, the intermediation approach is adopted and deposits are considered as inputs rather than outputs. Therefore, outputs are specified as

1. **Total loans** (Y1), which includes short term loans, trade bills and bills discounted, medium and long term loans and other loans, but excludes the loan loss reserves;
2. **Other earning assets** (Y2) such as short term and long term investment, deposits with central bank and other banks;
3. **Non-interest income** (Y3) derived from net fee and commission and other operating income.

Inputs are defined as

1. **Deposits** (I1) including short term deposits from customers and company, long term deposits, and other short term and long term borrowing and funding;
2. **Labor** (I2) specified as the registered full time employees in banks and
3. **Physical capital** (I3), often termed as fixed assets that provide the essential materials for bank running.

Correspondingly, input prices also have three parts. In banking efficiency literature, it is common to measure these input prices using ratios (see Altunbaş *et al.*, 2001 and 2000,

Table 3.4: Description of banking costs, outputs and input prices for 10 Asian countries at sample mean values for period 1998 to 2005 (mil US$)

	China	Hong Kong	India	Indonesia	Malaysia	Philippines	Singapore	Korea	Taiwan	Thailand
Total Costs	8151.09	888.755	2303.73	913.853	630.711	668.564	1330.26	2763.18	1025.95	1459.67
Total Assets	250663	27016.5	30589.1	9449.32	16172.4	9350.24	52316.5	52265.1	25614.3	34131.6
Equity Capital	11413	2007.23	1601.62	761.013	1360.11	1145.22	6043.95	2661.48	1738.06	2258.73
Output										
Y1: Loans	143497	12106.3	13321.9	2939.52	9663.18	4148.29	24363.7	33157.8	16522.4	21352.1
Y2: Other earning assets	90919.1	12075.3	13653.6	5314.6	5742.52	3513.39	22957.5	14142.3	6856.98	9627.02
Y3: Non-interest Income	718.099	257.540	479.537	130.938	160.395	176.397	405.094	392.399	133.650	302.608
Input Prices										
W1: Price of funds	0.02339	0.0352	0.0669	0.0923	0.0325	0.0535	0.0240	0.0508	0.0318	0.0345
W2: Price of labour	0.0310	0.0474	0.0363	0.0314	0.0565	0.0455	0.0427	0.0965	0.0511	0.0467
W3: Price of fixed assets	0.7913	0.4132	0.8566	1.1149	1.40161	0.9829	0.5470	0.4752	0.6662	0.0077
Number of Banks	48	22	56	35	21	23	3	16	41	15
Number of Obs.	277	154	382	230	149	157	22	115	301	103
Total Obs.	1890									

Source: Bankscope Database, 2007

Table 3.5: Summary statistics on cost, output quantities and input prices (1998-05) (mil US$)

		TC	Y1	Y2	Y3	W1	W2	W3
China	Mean	8151.09	143496.56	90919.05	718.10	0.02	0.03	0.79
	Median	674.89	8560.03	7975.85	76.60	0.02	0.02	0.72
	Min	17.11	165.84	74.59	1.29	0.01	0.00	0.20
	Max	89223.12	1711248.97	1360515.88	15445.99	0.07	0.13	2.99
	Stdev	18788.97	329716.99	210465.78	2022.08	0.01	0.03	0.40
	Mean	888.75	12106.29	12075.35	257.54	0.04	0.05	0.41
Hong Kong	Median	235.56	3787.93	3132.59	49.06	0.03	0.04	0.23
	Min	2.85	6.81	7.48	0.43	0.00	0.03	0.02
	Max	12029.85	172404.92	188905.52	5574.60	0.11	0.15	2.11
	Stdev	2002.77	25075.07	28334.41	713.20	0.02	0.02	0.45
India	Mean	2303.73	13321.92	13653.56	479.54	0.07	0.04	0.86
	Median	1245.04	6679.33	7094.85	252.93	0.07	0.03	0.74
	Min	26.70	2.62	31.07	1.37	0.02	0.01	0.04
	Max	33096.24	223272.78	242287.82	8559.59	0.13	0.55	5.08
	Stdev	4237.57	23427.69	28330.22	898.08	0.02	0.04	0.53
Indonesia	Mean	913.85	2939.52	5314.60	130.94	0.09	0.03	1.11
	Median	146.96	659.69	503.77	16.00	0.08	0.02	0.94
	Min	11.21	3.12	6.18	0.21	0.02	0.01	0.13
	Max	11711.25	28919.72	86683.13	2952.17	0.45	0.16	3.41
	Stdev	1805.73	5220.89	12343.56	318.94	0.07	0.03	0.65
Korea	Mean	630.71	9663.18	5742.52	160.40	0.03	0.06	1.40
	Median	492.18	7475.39	3726.90	108.71	0.03	0.04	0.98
	Min	10.23	49.85	124.11	0.00	0.01	0.01	0.25
	Max	3858.80	68453.82	31302.19	1364.49	0.12	0.27	5.77
	Stdev	681.58	11521.57	6331.62	197.24	0.02	0.05	1.15
Malaysia	Mean	668.56	4148.29	3513.39	176.40	0.05	0.05	0.98
	Median	346.46	1950.54	1933.85	87.78	0.05	0.04	0.75
	Min	23.82	108.33	49.81	1.64	0.02	0.01	0.15
	Max	2767.02	18357.31	17174.06	778.96	0.14	0.12	4.01
	Stdev	698.45	4556.51	4056.34	201.02	0.02	0.02	0.72
Philippines	Mean	1330.26	24363.71	22957.50	405.09	0.02	0.04	0.55
	Median	1256.03	26258.21	19871.07	337.58	0.02	0.04	0.43
	Min	870.76	11175.41	9898.19	74.01	0.01	0.02	0.30
	Max	2076.51	35787.53	46838.62	1163.40	0.05	0.06	1.02
	Stdev	344.92	7908.86	11222.55	267.36	0.01	0.01	0.24
Singapore	Mean	2763.18	33157.77	14142.27	392.40	0.05	0.10	0.48
	Median	2361.14	20014.30	11901.31	202.88	0.05	0.10	0.34
	Min	126.13	590.43	391.83	0.13	0.02	0.03	0.04
	Max	11499.07	180296.51	53672.62	2595.55	0.11	0.20	4.86
	Stdev	2505.98	37302.21	11732.64	457.57	0.02	0.04	0.56
Taiwan	Mean	1025.95	16522.37	6856.98	133.65	0.03	0.05	0.67
	Median	627.34	8947.50	2803.85	60.79	0.03	0.05	0.56
	Min	35.48	327.44	278.66	0.00	0.01	0.03	0.05
	Max	6142.07	79285.55	59212.72	1352.72	0.08	0.14	4.56
	Stdev	1035.85	16598.25	9401.47	191.61	0.02	0.02	0.53
Thailand	Mean	1459.67	21352.12	9627.02	302.61	0.03	0.05	0.01
	Median	1137.49	16956.97	4122.00	154.35	0.03	0.04	0.01
	Min	64.69	267.52	338.45	0.56	0.01	0.02	0.00
	Max	6399.89	68541.39	43980.29	2023.04	0.16	0.16	0.02
	Stdev	1346.15	19456.79	11530.05	350.43	0.02	0.03	0.00

Notes: TC=total costs; Y1=total loans; Y2=other earning assets; Y3=non-interest income; W1=price of funds; W2=price of labor; W3=price of fixed assets

Berger *et al.*, 2009 and many other studies listed in Table 2.1). The definition and calculation of these input prices are summarized as follows.

1. **Price of funding/deposit** (W1). It can be calculated by the ratio of interest expenses to total deposits.
2. **Price of labour** (W2). It can be calculated by the ratio of personnel expenses to number of employees. Collecting this variable is the hardest part of the data processing due to data on both the personnel expenses and number of employees. Not all the banks' reports collected from Bankscope provide the data for their personnel expenses or number of employees and sometime both of them are missing in the reports. This perhaps explains partly why limited cost banking efficiency studies in Asian countries exist in the literature. Therefore, in this context, this study may be the pioneering and the most comprehensive one in the area. The ways to handle out this problem rely on some reasonable and acceptable assumptions. For those banks missing the data for number of employees (i.e. banks in Hong Kong), I check the central bank website, the statistic department, annual reports each year, and other possible sources for this information. In circumstances of having data for some years but not all the periods, it is reasonable to assume that the growth rate of the number of employees is the same as the growth rate of total assets for a given bank (Altunbaş *et al.*, 2001, Vander Vannet, 2002 and Fu and Heffernan, 2007). For those banks missing the data for personnel expenses, the proxy such as the price of labour is used. If collectable, average wage rate in financial sector can be served as a good proxy for price of labour. If average wage rate is only available for a few years, I assume that the growth rate of wages is the same as the growth rate of total assets for finance sector. For those banks lacking of both of the data, they are left out of the sample.
3. **Price of fixed assets** (W3), defined as the ratio of other operating expenses (including depreciations and other capital expenses) to the fixed assets.
4.

3.2.2 Environmental Variables

Impact of cross-country environmental variables on banking efficiency analysis has been well discussed in this thesis. In the existing literature, two categories of environmental variables are usually considered:

(i) those that describe the main macroeconomic conditions, which determine the banking product demand characteristics, and

(ii) variables that describe the structure of the banking industry

Table 3.6: Description of the environmental variables using in the analysis

Environmental Variables	Description	Data Source
Macro Economic Indicators		
Z1: Density of Population	Population per square kilometres	IMF
Z2: GDP per capita	Ratio of GDP over inhabitants	As above
Z3: Inflation	Average CPI (Consumer Price Index). The base year is 2000	As above
Z4: Unemployment Rate (%)	Ratio of unoccupied inhabitants over the total population	As above, China Statistics Year Book (1998-2006), CIA- The World Fact book
Banking Structure		
Z5: Concentration	Assets of three largest banks as a share of assets of all commercial banks	World Bank (Database on the Financial Development and Structure)
Z6: Net Interest Margin	Accounting value of bank's net interest revenue as a share of its interest-bearing (total earning) assets	As above
Z7: Average Capital ratio (%)	Ratio of equity to total assets	Individual country's central bank website
Z8: Intermediation ratio	Ratio of total value of loans over total value of deposits	Individual country's central bank website

These two categories of environmental variables are described in Table 3.6. The first group is termed as macro-economic conditions that include a measure of population density, per capita income, inflation and unemployment ratio. These variables simply present the main macro conditions under which banks are providing their services.

1. **Density of population** (Z1). It is termed as population per square kilometres. Intuitively, banks may face a high cost when providing services in the area with a low population density, resulting in the potential low cost efficiency level.
2. **Per capita income** (Z2, ratio of GDP over inhabitants). Countries with high per capita income may have a banking system that operates in a mature environment, which can set a more competitive interest rate and profit margin. Thus the expectation for GDP per capita can be either positive, suggesting that the more

developed economy is, the higher costs banks incur as if banks are operating in a powerful and booming condition, they offer competitive interest rates and at the meanwhile the labour expenses may also be higher than before; or negative since the more developed economy, the higher possibility that their banking system may experience technological improvement and that individual customer may use a wide range of banking services, which may save banks' cost to search new customers and expand their services.

3. **Inflation** (Z3), measured as average CPI (Consumer Price Index) for each year with the base year 2000, is another important economic factor that may also influence the macro economic condition and financial system. The higher inflation may cause the depreciation of the national currency and the increasing of the products' price. To meet the basic living standard, individuals have to reduce their savings in the bank and producers may also have the same response for the demand of production. In the other end, number of loaners will increase due to the decreasing of cost of borrowing. Therefore, to fill the gap, banks have to increase their interest rates and look for the alternative funding source, which may increase banks' costs since deposits from individuals and companies are cheaper funding sources. Other sources like IPOs, issuing new shares in the market and rights issuing are far more expensive. So the higher costs are expected to be associated with higher inflation. As first indicated by Friedman (1956) and subsequently many other authors, the actual rate of inflation may be an indicator of expected rate of inflation. The expected rate of inflation is a key variable in determining the demand for money and hence affects the choice between holding wealth in the form of bank deposits relating to some other forms of wealth holding.
4. **Unemployment rate** (Z4). Higher unemployment rate may reflect the depression of the whole economy, which can spread to the financial sector. It may incur higher costs of banks' business and operation. Unemployment and job search expectations may also affect the precautionary demand for money and hence the level of deposits.

Four variables in the second category are used to reflect banking and financial structure.

1. **Bank concentration** (Z5). It is termed as a share ratio of total assets of three largest banks to those of all commercial banks. As an indicator of market power,

higher concentration may be associated with either higher costs or lower costs. As argued by Dietsch and Lozano-Vivas (2000), if higher concentration comes from the market power, costs will go with the same direction. However, cost may change in an opposite direction if higher concentration is caused by the superior management or superb efficiency of the production processes.

2. **Net interest margin** (Z6). It is calculated as the accounting value of bank's net interest revenue as a share of its total earning assets. It captures the difference between different banking industries in terms of their ability to convert deposits to loans. The better the ability, the higher net interest revenue and the lower costs of the banking industry.
3. **Capital ratio** (Z7), measured as ratio of equity capital to total assets. The higher capital ratio, the lower insolvency risk of the banking system.
4. **Intermediation ratio** (Z8). It captures the ability of banks to convert its deposits to loans. The higher the intermediation ratio, the lower the banks' cost.

Table 3.7 reports the average value of these environmental variables over 1998-2005 periods for these ten Asian countries. These arithmetic means suggests the large difference in the main conditions of banking activities across countries. A biased cost efficiency score may be expected without considering the influences of these variables.

Table 3.7: Summarized values of environmental variables in the Asian countries

	China	Hong Kong	India	Indon-esia	Korea	Mala-ysia	Philip-pines	Singap-ore	Taiwan	Thailand
Main Economic Indicator										
Z1	133.8	6075.3	351.4	110.2	476.0	73.8	261.5	5822.7	623.6	123.6
Z2	0.005	0.028	0.003	0.003	0.018	0.010	0.004	0.025	0.024	0.007
Z3	102.3	98.3	108.5	122.5	106.8	102.8	106.7	100.6	100.2	103.6
Z4	0.038	0.061	0.070	0.089	0.044	0.034	0.100	0.031	0.040	0.021
Banking Structure										
Z5	0.710	0.716	0.356	0.608	0.497	0.443	0.573	0.939	0.296	0.511
Z6	0.021	0.034	0.030	0.054	0.026	0.026	0.047	0.020	0.023	0.025
Z7	0.045	0.098	0.120	0.287	0.107	0.130	0.132	0.085	0.106	0.133
Z8	0.773	0.707	0.557	0.493	0.905	0.852	0.824	0.885	0.778	0.902

Source: IMF, World Economic Outlook Database, April 2007, World Bank (Database on Financial Development and Structure)

Chapter 4 Cost efficiency analysis in ten Asian banking industries[6]

4.1 Introduction

In the last two decades, the efficiency and productivity analysis of the banking industry has been investigated extensively. The emphasis of this area now spreads widely from scale and scope economies to the cost and profit efficiency. These studies apply the non-parametric frontier approach (e.g. DEA) and parametric frontier approach (e.g. SFA) to discuss the various issues in the efficiency measurement, for example issues of informing government policy like deregulation, merger and acquisition, problem loans and managerial performance, as well as addressing methodology issues and international comparisons.

Like most studies that focus on the banking industry of a specific country, most studies

[6] Two versions of this chapter has been published by Taylor & Francis:
Shen, Z., Liao, H. and Weyman-Jones, T. 2009, "Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions", *Journal of Chinese Economic and Business Studies*, vol. 7, no. 2, pp 199-218.
Shen, Z., Liao, H. and Weyman-Jones, T. 2009, "Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions" in *China's Three Decades of Economic Reforms*, eds. X.H. Liu and W. Zhang, Routledge, Taylor & Francis Group, UK.

of international comparison employ the efficiency measurement in the developed US and European market. Only one cross-country study can be found in the literature and it measures cost efficiency of 111 commercial banks in four South Asian banking markets (Perera *et al.*, 2007). However, there is no efficiency studies investigating and comparing banks' performance of the major Asian economies. The potential reason of this literature gap may attribute to the lack of the satisfactory quality data, development of the whole economy and financial system. Therefore, this thesis will be one of the first to address this issue.

As suggested in the existing literature of cross-country studies, due to different geographical and macro economic conditions, countries differ from each other substantially. As a result, differences in the managerial ability of banks may not be the only way to explain the observed differences in banking performance. Therefore, in setting up the efficient frontier, cross-country heterogeneous factors cannot be excluded. This study, along with some other current ones (Casu and Girardone, 2004; Dietsch and Lozano-Vivas, 2000; Fries and Taci, 2004; Kasman and Yildirim, 2006; Lozano-Vivas *et al.*, 2002), takes account of the impact of the cross-country heterogeneities banks' production technology. Efficiency results are compared with and without incorporation of cross-country environmental variables.

Moreover, from the estimating perspective, this study is among the first to adopt the panel data stochastic frontier approach with an attempt to measure cost efficiency. Efficiency scores in the previous cross-country studies are usually estimated using a cross-sectional stochastic frontier approach, which suffers from shortcomings that can be solved by using panel data framework, as suggested by Schmidt and Sickles (1984).

The structure of this chapter is organized as follows. Section 4.2 will review the existing cross-country studies that shed light on the motivation of this study. In Section 4.3 panel data stochastic frontier methodology will be introduced from the cost perspective, with model specifications discussed. Section 4.4 presents the empirical results and section 4.5 provides further discussions on two long-standing debates. Section 4.6 concludes.

4.2 Literature of cross-country banking efficiency studies

A review of cross-country studies was first carried out by Berger and Humphrey (1997). Six cross-country comparisons are reviewed, with five based on DEA and one based on the parametric approaches (distribution free approach, DFA and thick frontier approach, TFA). No other dedicated survey of cross-country efficiency studies can be found. Therefore, to shed light on my own interests, a survey of 15 cross-country studies is conducted (see Table 4.1).

4.2.1 Applied countries of cross-country studies

In the 1990s, two studies measured and compared the technical efficiency of banking firms. One concentrated on three Nordic countries and the other focused on 8 EU countries. Twelve other cross-country studies examined cost efficiency and profit efficiency, suggesting that the heart of the efficiency analysis has now shifted to cost and profit efficiency, which indicate that researchers should evaluate the managerial performance of a banking firm from the cost and profit perspectives. Nine of them made a comparison of cost efficiency among developed European countries, U.S and Japan. Two other studies measured the banking performance in transition countries and one in Latin American countries. There is only one Asian efficiency study that measures cost efficiency in four South Asian banking markets, Bangladesh, India, Pakistan and Sri Lanka, over 1997-2004. Thus, this work intends to fill the literature gap of no cross-country comparisons among the major Asian banking industries.

4.2.2 Utilized measurement technology

In order to compare the efficiency at an international level, existing cross-country studies coincidently build up a common frontier (either production or cost or profit frontier) as a benchmark. Technical efficiency (cost or profit efficiency) is then calculated or estimated by using frontier approaches such as DEA, SFA and DFA. In my survey, five studies utilize DEA to measure the efficiency, while nine studies use SFA and two adopt DFA. DEA is a linear programming technique where observed banking firms are used to form the efficient frontier as the piecewise linear combinations that

Table 4.1: Survey of cross-country studies

No.	Studies	Key features	Efficiency score	General conclusion
1	Allen and Rai (1996)	- Cost and profit efficiency - 15 EU countries and US from 1988-92 - SFA and DFA - Intermediation approach	Average cost efficiency: 0.78/0.85 (Large/small banks)	Large banks was found to be the most inefficient as well as diseconomies of scale.
2	Altunbaş *et al.* (2001)	- Cost efficiency - 15 EU countries from 1989-97 - SFA - Intermediation approach	Average cost efficiency: 0.75 to 0.8 across different sizes	Efficiency scores vary across country, bank sizes and over time, suggesting that great cost savings can be achieved by improving managerial ability.
3	Berg *et al.* (1993)	- Technical efficiency - 3 Nordic Countries - DEA - Value-added approach	Average technical efficiency: VRS: Fin: 0.53, Nor: 0.58, Sw: 0.78 CRS: Fin: 05, Nor: 0.41, Sw: 0.69	By setting up the common frontier, Sweden is found to be the most efficient country among three.
4	Casu and Girardone (2004)	- Cost and profit efficiency - 5 EU countries from 1993-97 - DEA and SFA - Intermediation approach - Country-specific environmental variables	Average cost efficiency: 0.855 Average Profit efficiency: 0.54 in 1993 to 0.9 in 1997	Efficiency level is not converged despite harmonization of EU. Country factors appear to be important in explaining inefficiency. Profit efficiency shows a sharp growth over time.
5	Carvallo and Kasman (2005)	- Cost efficiency - 16 Latin American countries from 1995-99 - SFA - Dual approach	Average cost efficiency: 0.822	The efficiency level varies across countries dramatically while the largest economy appears to be the most efficient.
6	Dietsch and Lozano-Vivas (2000)	- Cost efficiency - Spain and France from 1988-92 - DFA - Value-added approach - Country-specific environmental variables	Average cost efficiency: Without/with environmental variables Fr: 0.58/0.89, Sp: 0.09/0.75	Cross-country heterogeneous factors have an important role on explaining part of the inefficiency. Without considering them, the efficiency estimates would be biased.
7	Fries and Taci (2004)	- Cost efficiency - 15 transition countries from 1994-01 - SFA and DFA - Production approach - Country-specific environmental variables	Average cost efficiency: Without/with environmental variables 0.63/0.71	Privatization increases the operating efficiency since private banks are more efficient than state-owned banks. Foreign ownership does help improve the managerial ability.
8	Kasman and Yildirim (2006)	- Cost and profit efficiency - 8 central and eastern European countries from 1995-02 - SFA - Dual approach - Country-specific environmental variables	Average cost efficiency: 0.793 Average profit efficiency: 0.633	Cost and profit efficiency spreads substantially among different countries. Foreign banks perform more efficient than domestic banks.

Table 4.1: Survey of cross-country studies (continued)

No.	Studies	Key features	Efficiency score	General conclusion
9	Lozano-Vivas *et al.* (2002)	- Cost efficiency - 10 EU countries in 1993 - DEA - Value-added approach - Country-specific environmental variables	Average cost efficiency: Without/with environmental variables Bel: 0.42/0.79, Den: 0.2/0.75, Fr: 0.24/0.41, Ger: 0.27/0.58, It: 0.25/0.33, Lux: 0.49/0.62, Net: 0.37/0.52, Por: 0.16/0.8, Sp: 0.19/0.82, UK: 0.22/0.59	Country-specific environmental variables exercise a strong influence over the behavior of each country's banking industry
10	Maggi and Rossi (2003)	- Cost efficiency - 15 EU countries and US from 1995-98 - DFA - Production and value-added approach	Mean cost efficiency: EU: 0.64, US: 0.62	Fourier flexible functional form fit the data better than the translog functional form.
11	Maudos *et al.* (2002)	- Cost and profit efficiency - 10 EU countries from 1993-96 - SFA and DFA - Intermediation approach	Average cost efficiency: 0.83 (DFA), 0.77 (FE), 0.84 (RE) Average profit efficiency: 0.45 (DFA), 0.22 (FE), 0.52 (RE)	Profit efficiency is lower than cost efficiency implying that the most important inefficiency on the revenue side.
12	Pastor *et al.* (1997)	- Technical efficiency - 8 EU countries for the year 1992 - DEA - Value-added approach	Average technical efficiency: 0.86	Technical efficiency varies among countries.
13	Perera *et al.* (2007)	- Cost efficiency - 4 South Asian countries over 1997-2004 - SFA - Two outputs and two input prices - Cross-country environmental variables	Average efficiency: Across the sample: 0.89 Bangladesh: 0.89, India: 0.92 Pakistan: 0.89, Sri Lanka: 0.87	Overall efficiency of South Asian banks declined over 1997–2004. Larger banks and banks with widespread ownership through stock exchange listings were found to be more cost efficient. In contrast, state-owned banks were less efficient.
14	Vander Vennet (2002)	- Cost and profit efficiency - 17 EU countries from 1995-96 - SFA - Intermediation approach	Average cost efficiency: Universal: 0.8, specialized: 0.7	Universal banks perform more efficiently than separated banks.
15	Weill (2004)	- Cost efficiency - 5 EU countries from 1992-98 - DEA, SFA, DFA - Intermediation approach	Average cost efficiency: Fr: 0.71, Ger: 0.83, It: 0.84, Sp: 0.78, Sw: 0.66	Lack of robustness between approaches, even if there are some similarity in particular between parametric approaches. Mixed evidences of the relationship between bank size and efficiency were found.

connect the best practice observations, yielding the convex production possibilities set. Those banks that lie on the frontier are the most efficient. On the other hand, banks that do not lie on that surface are considered as inefficient. Since DEA suffers a key drawback of not allowing the random error, SFA is preferred. However, these studies use a cross-sectional stochastic frontier, which suffers from shortcomings such as inconsistent estimates of inefficiency, no time-varying inefficiency, strong distributional assumption on the inefficiency term, uncorrelatedness of inefficiency with the regressors (see p.69-70). These shortcomings can be solved when panel data stochastic frontier approaches are adopted.

4.2.3 Output and input specifications

The difference in identifying the role of deposits (see p.14 for detailed explanation) is also reflected in cross-country studies since in our survey, six studies adopt the intermediation approach, two adopt the production approach and dual approach respectively and five apply the value-added approach. However, although it is important to acknowledge the dual role of financial institutions and deposits, to define deposits as both output and input also depends on the availability of data on deposits. It is reasonable to consider deposits as output because financial institutions provide transactions and document processing services for customers. Thus, as suggested by production approach and dual approach, the number of deposits accounts rather than the value of deposits should be considered as output since these documenting services are provided to any depositors no matter of the sizes of their deposits accounts. This may not be a problem if DEA is used but it may be a problem when parametric approach such as SFA is adopted. This is because based on test results of statistical inferences and theoretical properties of using cost function on my data set and two other existing studies, I find that when only data on the value of deposits are available, it is not suitable to consider deposits as output since this manner fails the tests, as suggested in section 4.5. However, regarding to my literature survey in chapter 2, since majority banking efficiency studies use the value of total deposits rather than the number of deposits accounts, it is more suitable to use intermediation approach and consider deposits as an input.

4.2.4 Cross-country heterogeneous factors

As argued by Berger and Humphrey (1997), *"cross-country comparisons are difficult to interpret because the regulatory and economic environments faced by financial institutions are likely to differ importantly across nations…Such cross-country differences were not specified when a 'common' frontier was being estimated and this may affect the cross-country results."* Nine studies in my survey share the assumption that banks of those countries in comparison provide banking services under the same production process and conditions. Therefore, the observed inefficiency in these studies is attributed to poor managerial performance. However, countries may differ not only geographically, but also by the macro-economic power and financial regulatory requirement. Differences in managerial abilities between banks may not be the only reason for observed difference in banking performance. Therefore, in setting up the common frontier, cross-country heterogeneous factors cannot be excluded. Six recent cross-country studies introduce such heterogeneous factors by including environmental variables reflecting the various differences amongst countries. The empirical results show that efficiency scores are higher when cross-country heterogeneity is considered, indicating that cross-country differences may explain part of the inefficiency estimated when these factors are excluded and neglecting these factors may cause efficiency scores to be underestimated.

4.3 Methodology

4.3.1 Panel Data Stochastic Frontier Approach

First proposed by Aigner *et al.* (1977) and Meeusen and van den Broeck (1977), stochastic frontier approach has been widely used in the efficiency literature. These models allow for technical inefficiency, but they also acknowledge the fact that random shocks outside the control of producers can affect the output of the producer. By forming a composed error term, they separate the idiosyncratic errors from the technical inefficiency. Therefore, technical inefficiency would not be contaminated by the random noises that shouldn't be considered as inefficiency.

Rather than using a cross-sectional stochastic frontier models that has been widely adopted in the cross-country studies, panel data framework is adopted due to the limitations of cross-sectional models discussed in section 2.2. The methodological development of panel data stochastic frontier approaches has been discussed in section 2.2.4.2.2 based on the framework of production function. Since this work is to measure and compare cost efficiency in Asian banking industries, it is necessary to demonstrate how panel data stochastic frontier approaches are modelled under the framework of cost function.

The classical panel data cost frontier model[7] can be written as follows

$$\ln C_{it} = \alpha_0 + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} + u_{it} \qquad [4.1]$$

where C_{it} stands for total costs for firm i (i=1…N) at time t (t=1…T), y_{jit} and w_{mit} represents the jth (j=1…J) output and the input price for the mth (m=1…M) input for firm i at time t, respectively. $u_{it} \geq 0$ represents cost inefficiency while $v_{it} \sim \text{iid}\left(0, \sigma_v^2\right)$ stands for the random errors that are beyond the control of firms.

If cost efficiency is time-invariant, the fixed-effects (FE) and random-effects (RE) model can be adopted and [4.1] will be modified as

$$\ln C_{it} = \alpha_0 + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} + u_i \qquad [4.2]$$

The FE model assumes that v_{it} is uncorrelated with the regressors. No distributional assumption is made on u_i and it can be correlated with the regressors or v_{it}. Since u_i is treated as fixed, it becomes the producer specific intercept to be estimated with β_j and δ_m by using the least squares with dummy variables (LSDV for short). The model [4.2] will be modified as

$$\ln C_{it} = \alpha_{0i} + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} \qquad [4.3]$$

[7] For simplicity, Cobb-Douglas functional form is used here.

where $\alpha_{0i} = \alpha_0 + u_i$. By using the transformation $\hat{\alpha}_o = \min(\hat{\alpha}_{0i})$ and $\hat{u}_i = \hat{\alpha}_{0i} - \hat{\alpha}_0$, cost efficiency can be obtained from $CE_i = \exp(-\hat{u}_i)$. However, the FE model suffers a potential defect that inefficiency u_i will capture all the time-invariant effects that vary across firms, including time-invariant inefficiency and other time-invariant factors such as the regulatory environment. Therefore, the cost inefficiency might be overestimated. Schmidt and Sickles (1984) argued that this problem could be fixed with RE model since the assumption of cost inefficiency to be randomly distributed with constant mean and variance and not correlated with any regressors and v_{it} could allow some time-invariant regressors in the model. The RE model can be written as

$$\begin{aligned} \ln C_{it} &= [\alpha_0 + E(u_i)] + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} + [u_i - E(u_i)] \\ &= \alpha_0^* + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} + u_i^* \end{aligned}$$

[4.4]

where $\alpha_0^* = \alpha_0 + E(u_i)$ and $u_i^* = u_i - E(u_i)$. $E(u_i)$ is the mean of cost inefficiency. Cost efficiency can be estimated by using generalized least squares (GLS). Pitt and Lee (1981) extends this RE model by adding further distributional assumption on v_{it} and u_i, allowing $v_{it} \sim \text{iid } N(0, \sigma_v^2)$ and $u_i \sim \text{iid } N(0, \sigma_u^2)$. Then maximum likelihood estimation (MLE) can be used to estimate cost efficiency, and MLE is consistent and asymptotically efficient.

However, it is not so convincing to expect cost inefficiency to be time-invariant in a long time period, especially when the open environment is competitive. The longer is the time period the more desirable it is to relax this time-invariant assumption. In the literature, a number of studies have adopted Battese and Coelli (1992, 1995) model, which try to relax the assumption of time-invariant inefficiency by introducing the additional term $u_{it} = \exp(-\eta(t-T)) \cdot u_i$ into Pitt and Lee (1984) RE model. Cost inefficiency is said to decrease in an increasing rate if $\eta > 0$ and increase in an increasing rate if $\eta < 0$.

However, as argued in Greene (2004, 2005), the above classical panel data models have one limitation of not including the time-invariant heterogeneities in those models. And if these time-invariant heterogeneities do exist but are not included, all these heterogeneities will be pushed into the intercept α_0 and finally into inefficiency u_i, resulting in an overestimated cost inefficiency score. This limitation is relaxed by introducing so called 'true' fixed-effects and random-effects models. In Greene's 'true' FE model, firm specific constant terms are introduced in the stochastic frontier models, written as,

$$\ln C_{it} = \alpha_i + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} + u_{it}$$

[4.5]

where α_i incorporates all the time-invariant firm specific heterogeneities and the regressors, random errors and inefficiency terms are mutually uncorrelated and u_{it} is not restricted to be time-invariant. But one has to sacrifice the freedom of no distributional assumption on random noises and inefficiency. Then the MLE can be used to estimate cost inefficiency. The 'true' RE model uses the random constant term to embody the time-invariant firm specific heterogeneities in the cost function, written as:

$$\ln C_{it} = (\alpha + w_i) + \sum_j \beta_j \ln y_{jit} + \sum_m \delta_m \ln w_{mit} + v_{it} + u_{it}$$

[4.6]

However, these 'true' SFA models may overcompensate for time-invariant heterogeneities since the inefficiency can be time-invariant to some extent in financial systems where performance related incentives are weak or absent. If there is persistent inefficiency, it is completely absorbed in the firm specific constant term that also captures all the time-invariant heterogeneities. Consequently, as the classical panel data model might overestimate the inefficiency, the 'true' SFA models might underestimate it.

Therefore, in a cross-country banking efficiency comparison, with the availability of the information on cross-country heterogeneous factors in geographic, economic and regulatory perspectives, one should measure time-invariant effects as both time-invariant heterogeneities and time-invariant inefficiency. Obviously, the classical FE model cannot include cross-country heterogeneities but only classical RE model can be used for this purpose. This chapter will first compare the efficiency estimates from the classical panel data models with the ones from Greene's 'true' SFA models. Then by incorporating cross-country environmental variables in the model, whether they are important sources to explain banks' performance will be examined.

4.3.2 Model specification

In this chapter, translog cost functional form is adopted and can be written as:

$$
\begin{aligned}
\ln C_{it} &= \alpha + \sum_{j=1}^{3} \beta_j \ln y_{jit} + \sum_{m=1}^{3} \delta_m \ln w_{mit} + \phi_i \ln E_{it} \\
&+ \frac{1}{2}\sum_{j=1}^{3}\sum_{k=1}^{3} \beta_{jk} \ln y_{jit} \ln y_{kit} + \frac{1}{2}\sum_{m=1}^{3}\sum_{n=1}^{3} \delta_{mn} \ln w_{mit} \ln w_{mit} + \sum_{j=1}^{3}\sum_{m=1}^{3} \gamma_{jm} \ln y_{jit} \ln w_{mit} \\
&+ \theta_1 t + \frac{1}{2}\theta_2 t^2 + \sum_{j=1}^{3} \xi_j \ln y_{jit} t + \sum_{m=1}^{3} \zeta_m \ln w_{mit} t + \sum_{l=1}^{8} \tau_l Z_{lit} + v_{it} + u_{it}
\end{aligned}
\tag{4.7}
$$

where total costs of bank i $C_{it}(i = 1,..., N)$, observed for t times, are given as a function of three outputs $y_{jit}(j = 1,...,3)$, three input prices $w_{mit}(m = 1,...,3)$, the equity capital ratio E_{it} and environmental variables Z_{lit} that account for cross-country heterogeneities that vary over time[8] but not across banks within each country, but they do vary across countries. Equity capital ratio is considered as a risk control factor since equity capital may influence the probability of banks' failure and so interest costs. Also, a bank's capital level will directly affect costs by providing an alternative funding source. Data descriptions are provided in chapter 3.

To ensure cost efficiency estimates are truly estimated from the cost function, following properties of cost function suggested by McFadden (1978:10) and Kumbhakar and Lovell (2000:34) should be satisfied:

(i) non-decreasing in **y**, as $\partial \ln C_{it} / \partial \ln y_{jit} \geq 0$

[8]Although cross-country heterogeneities are time-varying over sample period, however, these differences are trivial and one cannot rule out time-invariant characteristics in these time-varying heterogeneities. Moreover, based on the LR test, it is suggested that treating these time-varying heterogeneities as time-invariant, estimates of the parameters and inefficiency scores are quite similar, which suggest that time-invariant heterogeneities are the latent characteristics in cross-country heterogeneities.

(ii) non-decreasing in **w**, as $\partial \ln C_{it} / \partial \ln w_{mit} \geq 0$

(iii) homogenous of degree one in **w**, as $C_{it}\left(\mathbf{y}, \frac{\mathbf{w}}{w_{Mit}}, t\right) = \frac{C_{it}(\mathbf{y}, \mathbf{w}, t)}{w_{Mit}}$, written in log terms as $\ln C_{it}\left(\mathbf{y}, \frac{\mathbf{w}}{w_{Mit}}, t\right) = \ln C_{it}(\mathbf{y}, \mathbf{w}, t) - \ln w_{Mit}$

(iv) concave in **w**

(v) scale elasticity of cost function can be measured by:

$$E = \sum_{j=1}^{J} \partial \ln C_{it} / \partial \ln y_{jit} = \sum_{j=1}^{J} ey_j \qquad [4.8]$$

By imposing property (iii), the cost function for estimation can be written as

$$\begin{aligned}
\ln \frac{C_{it}}{w_3} &= \alpha + \sum_{j=1}^{3} \beta_j \ln y_{jit} + \sum_{m=1}^{2} \delta_m \ln \frac{w_{mit}}{w_3} + \phi_i \ln E_{it} \\
&+ \frac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \beta_{jk} \ln y_{jit} \ln y_{kit} + \frac{1}{2} \sum_{m=1}^{2} \sum_{n=1}^{2} \delta_{mn} \ln \frac{w_{mit}}{w_3} \ln \frac{w_{nit}}{w_3} + \sum_{j=1}^{3} \sum_{m=1}^{2} \gamma_{jm} \ln y_{jit} \ln \frac{w_{mit}}{w_3} \\
&+ \theta_1 t + \frac{1}{2} \theta_2 t^2 + \sum_{j=1}^{3} \xi_j \ln y_{jit} t + \sum_{m=1}^{2} \zeta_m \ln \frac{w_{mit}}{w_3} t + \sum_{l=1}^{8} \tau_l Z_{lit} + v_{it} + u_{it}
\end{aligned}$$

[4.9]

To ensure the symmetry condition, following restrictions $\beta_{jk} = \beta_{kj}$ and $\delta_{mn} = \delta_{nm}$ is imposed. Monotonicity properties (i) and (ii) are checked by calculating the elasticities of output and input prices, $ey_j = \partial \ln C_{it} / \partial \ln y_{jit}$ and $ew_m = \partial \ln C_{it} / \partial \ln w_{mit}$, which can be expressed in the terms of coefficients of the fitted cost function. Concavity condition of cost function in input prices **w** is satisfied when the Hessian matrix of cost function with respect to input prices **w** is negative semi-definite. It is derived as:

$$H(w) = \boldsymbol{\delta} - \hat{\mathbf{s}} + \mathbf{s}\mathbf{s}^T \qquad [4.10]$$

In [4.10], $\boldsymbol{\delta}$ is the matrix of second order coefficients of input prices in the cost function. $\mathbf{s}$ is the column matrix of share equations, $s_m = \partial \ln C_{it} / \partial \ln w_{mit} = ew_m$, also known as the shepherd lemma and $\mathbf{s}^T = [s_1, ..., s_M]^T$. $\hat{\mathbf{s}}$ is the diagonal matrix with the share s_m on the main diagonal.

4.4 Empirical Results

4.4.1 Cost efficiency estimates without considering cross-country heterogeneities

Table 4.2 presents the parameter estimates of panel data models without considering cross-country heterogeneities and average cost efficiency scores across the sample and for individual country and region are reported in Table 4.3. As expected, when not accounting for the impact of cross-country heterogeneities, lower efficiency scores are observed from classical panel data models than from Greene's true SFA model.

The average cost efficiency score is about 0.2718 from the FE model, which is less than the efficiency score from the RE model (0.4294 in the Pitt and Lee model and 0.4249 in Battese and Coelli model) for the whole sample data[9]. A similar result can be found when collecting cost efficiency score for specific country. This result coincides with our expectation for the following three reasons: First of all, the FE model assumes that inefficiency u_i is fixed, which eventually will absorb all the time-invariant effects that are time-invariant heterogeneities but not inefficiency. Second, the FE model assumes that inefficiency u_i can be correlated with the regressors while Pitt and Lee RE model assumes that inefficiency u_i is uncorrelated with the regressors and follow the half-normal distribution. The higher efficiency gain might be the payoff of the distributional assumption. Finally, the FE model use LSDV to estimate the parameters and the firm-specific intercept α_{0i}. By using the transformation $\hat{\alpha}_0 = \min(\hat{\alpha}_{0i})$ and $\hat{u}_i = \hat{\alpha}_{0i} - \hat{\alpha}_0$, cost efficiency can be obtained by $CE_i = \exp(-\hat{u}_i)$. Since these efficiency estimates are relative to the best performing bank in the sample, they are viewed as the relative efficiency rather than the absolute efficiency score. Therefore, by using the FE model, an upward bias may be imposed to inefficiency estimates as the best performed bank is labeled as 100% efficient, which may be only 90% efficient but outperform other banks.

---

9 Hausman test favours the FE than RE. However, the RE model can still be used. The Hausman test is based on the classical panel data models and concentrates on whether the individual-specific effects α_i is random- or fixed-effects. But in efficiency studies, what researchers are interested is whether inefficiency is random or fixed effects. The Hausman test would simply reject the former based on the existence of correlatedness between the individual-specific effects and the regressors, but not necessarily the inefficiency itself. Since inefficiency in the FE model might capture all the time-invariant heterogeneities, there is a possibility of the correlatedness between those heterogeneity and the regressors but ultimately imposing this correlatedness to inefficiency, which actually is not correlated with the regressors.

Table 4.2: Estimated parameter coefficients for classical and 'true' panel data models without considering cross-country heterogeneities

	Time-invariant models						Time-varying models			
	FE		Pitt and Lee RE		Battese and Coelli RE		True FE		True RE	
Parameters	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio
Constant			-0.555***	-14.353	-0.567***	-17.201			0.238***	185.042
Y1	0.444***	30.093	0.497***	39.619	0.504***	45.578	0.619***	14.420	0.533***	284.885
Y2	0.271***	17.075	0.308***	23.413	0.296***	23.430	0.399***	9.644	0.337***	197.215
Y3	0.052***	6.650	0.062***	9.909	0.050***	8.088	0.032	1.444	0.035***	34.125
W1	0.693***	55.795	0.701***	62.139	0.705***	67.767	0.893***	16.759	0.708***	518.191
W2	0.175***	14.639	0.176***	19.540	0.155***	18.462	0.029	0.980	0.178***	153.395
lnE	-0.070***	-6.854	-0.071***	-9.917	-0.068***	-9.981	-0.094***	-3.309	-0.062***	-73.811
Y11	0.050***	17.093	0.055***	25.715	0.059***	26.854	0.081***	9.219	0.060***	68.512
Y12	-0.114***	-13.997	-0.126***	-20.060	-0.132***	-22.774	-0.163***	-7.563	-0.130***	-67.370
Y13	0.012***	3.565	0.011***	5.830	0.001***	4.994	0.024**	2.325	0.010***	7.988
Y22	0.072***	16.235	0.080***	20.180	0.082***	20.361	0.149***	10.534	0.081***	70.710
Y23	-0.030***	-7.042	-0.032***	-12.528	-0.035***	-12.590	-0.115***	-8.332	-0.024 ***	-22.880
Y33	0.012***	6.641	0.014***	12.542	0.014***	13.048	0.034***	5.839	0.009***	17.033
W11	0.041***	8.378	0.031***	7.031	0.032***	7.010	-0.060***	-3.777	0.037***	45.475
W12	-0.062***	-7.378	-0.050***	-6.472	-0.051***	-6.389	0.114***	4.464	-0.045***	-36.754
W22	0.017***	3.658	0.017***	4.400	0.017***	4.234	-0.042***	-3.148	0.017 ***	24.372
Y1W1	-0.005	-0.771	-0.014***	-3.260	-0.016***	-4.303	-0.081***	-4.260	-0.013***	-9.955
Y1W2	0.001	0.171	0.008*	1.703	0.010**	2.213	0.025	1.489	-0.000	-0.147
Y2W1	0.045***	6.501	0.045***	8.396	0.044***	9.130	0.051**	2.438	0.034***	23.090

Table 4.2: Estimated parameter coefficients for classical and 'true' panel data models without considering cross-country heterogeneities (continued)

	Time-invariant models				Time-varying models					
	FE		Pitt and Lee RE		Battese and Coelli RE		True FE		True RE	
Parameters	**Coeff.**	**t-ratio**	**Coeff.**	**t-ratio**	**Coeff.**	**t-ratio**	**Coeff.**	**t-ratio**	**Coeff.**	**t-ratio**
Y2W2	-0.024***	-3.836	-0.026***	-4.945	-0.028***	-5.654	-0.046**	-2.427	-0.015***	-11.992
Y3W1	-0.010**	-2.299	-0.009***	-3.038	-0.006**	-2.059	0.032**	2.438	-0.003***	-2.957
Y3W2	0.001	0.381	0.000	-0.093	-0.002	-0.623	-0.002	-0.187	0.001*	1.887
T	0.005	1.248	-0.009***	-2.780	-0.059***	-14.037	0.009	0.974	-0.012***	-23.912
SQRT	0.001	0.908	-0.000	-0.013	-0.003***	-3.756	-0.014***	-6.519	-0.003***	-22.371
Y1T	0.005**	2.295	0.002	1.148	-0.000	-0.100	-0.011	-1.469	0.002***	4.223
Y2T	-0.008***	-3.024	-0.008***	-3.627	-0.009***	-4.502	-0.012	-1.567	-0.011***	-19.071
Y3T	0.003**	2.280	0.005***	4.879	0.003***	3.128	0.016***	3.395	0.004***	13.421
W1T	0.000	0.167	-0.003	-1.488	-0.014***	-6.163	-0.046***	-5.395	-0.001**	-2.507
W2T	0.000	0.064	0.004**	2.328	0.008***	4.090	0.060***	8.233	0.005***	13.521
R-square	0.998									
σ	2.133		0.962		1.137		0.501***	18.010	0.185***	134.786
η					-0.053***	-18.612				
λ			9.050***	3.328	11.097***	1304.822	0.988***	9.087	382.977	0.901
σ_u			0.956***	10.178	1.138***	8.301	0.352		0.185	
L.L.F.			873.945		921.210		-214.189		1096.379	

Note: ***, **, * denote the rejection of the null at the 1%, 5%, 10% level, respectively.

Table 4.3: Estimated average efficiency score from panel data models without incorporating the heterogeneous factors

	Time-invariant Models		Time-varying Models		
	FE	RE (Pitt & Lee)	B&C	True FE	True RE
Whole Sample Efficiency Score					
Mean (Standard Dev.)	0.2718 (0.1384)	0.4294 (0.1421)	0.4249 (0.1422)	0.7735 (0.0061)	0.9904 (0.0001)
Country Specific Efficiency Score					
China	0.2396 (0.0976)	0.3966 (0.1089)	0.3883 (0.1030)	0.7737 (0.0035)	0.9904 (0.0001)
Hong Kong SAR	0.3118 (0.1241)	0.4784 (0.1326)	0.4773 (0.1392)	0.7696 (0.0146)	0.9904 (0.0001)
India	0.2429 (0.0786)	0.4044 (0.0833)	0.3994 (0.0870)	0.7750 (0.0034)	0.9904 (0.0002)
Indonesia	0.3278 (0.1547)	0.4537 (0.1660)	0.4462 (0.1610)	0.7724 (0.0068)	0.9904 (0.0002)
Korea	0.2480 (0.0393)	0.4589 (0.0458)	0.4479 (0.0472)	0.7734 (0.0029)	0.9904 (0.0000)
Malaysia	0.4697 (0.2766)	0.6452 (0.2192)	0.6427 (0.2171)	0.7744 (0.0045)	0.9904 (0.0002)
Philippines	0.2668 (0.0931)	0.4036 (0.0984)	0.4086 (0.0982)	0.7754 (0.0017)	0.9904 (0.0000)
Singapore	0.2384 (0.0190)	0.4554 (0.0241)	0.4539 (0.0192)	0.7747 (0.0013)	0.9904 (0.0000)
Taiwan Province of China	0.2556 (0.0581)	0.4305 (0.0662)	0.4305 (0.0654)	0.7732 (0.0061)	0.9904 (0.0000)
Thailand	0.1024 (0.0280)	0.1978 (0.0371)	0.1878 (0.0437)	0.7712 (0.0059)	0.9904 (0.0002)

Table 4.4: Correlation matrix among panel data models without accommodating cross-country heterogeneities

	FE	RE(P&L)	BC	True FE	True RE
FE	1.00000	0.94845	0.91567	-0.01826	0.08267
RE(P&L)	0.94845	1.00000	0.96337	0.00004	0.05517
BC	0.91567	0.96337	1.00000	0.00961	0.05207
True FE	-0.01826	0.00004	0.00961	1.00000	0.09102
True RE	0.08267	0.05517	0.05207	0.09102	1.00000

Despite the above differences in model setup, presumed assumptions, FE and RE models are actually quite similar in generating efficiency estimates, based on the results of correlation matrix among the classical FE and RE models (see Table 4.4), in which the correlation between the pair of FE and RE estimates is 0.9485 (FE vs. Pitt and Lee RE) and 0.9157 (FE vs. Battese and Coelli RE). This result is consistent with the literatures comparing the use of FE and RE model. In Greene (2004)'s WHO paper, the author finds the correlation between the pair of estimates is almost one for time-invariant FE and RE model for both DALE and COMP measures of health care outcome. Moreover, as summarized in Kumbhakar and Lovell (2000: 106-107), Gong and Sickles (1989) find that FE and RE (both GLS and MLE) approaches generated the similar estimates of efficiency in terms of both correlation and rank correlation. Gathon and Perelman (1992) report the Spearman rank correlation above 0.80 when comparing three approaches using European railway data. While using the U.S. banking data, Bauer *et al.* (1993) find the similar correlation about 0.89 between FE and RE model based on regression but notable difference between these and Pitt and Lee approach. More similar findings are provided in Bauer and Hancock's (1993) U.S. Federal Reserve check processing facilities data and Ahmad and Bravo-Ureta's (1996) U.S. dairy farm data.

Compared with the classical FE and RE models, average cost efficiency scores from Greene's 'true' FE and RE models are very high, at 0.7735 and 0.9904 respectively. This is expected since the construction of the 'true' SFA model enables us to move some of the time-invariant heterogeneities out of the inefficiency term and lower inefficiency will be observed. This is consistent with the estimates of $\lambda = \sigma_u / \sigma_v$, where for classical FE and RE model, λ are 9.050 and 11.097, indicating the variance decomposition is dominated by the inefficiency term and leading to higher inefficiency level. However, for 'true' FE and RE model, λ are 0.988 and 0[10], indicating relatively lower inefficiency level. The difference in model specifications between classical and 'true' panel data models is also reflected in Table 4.4, where the correlations between the pair of classical and 'true' FE and RE models are less than 0.08, indicating the existence of large volume of time-invariant heterogeneities. It is not surprising as my banking data consists of ten major Asian banking industries that differ in the aspects of geography, culture, economy power and banking and financial structure. Therefore, it is very important to take account these heterogeneities in the measurement of banking performance. However, the fact that cost efficiency scores are so high and evenly distributed across countries and regions, especially from the 'true' random effects model, is not expected but unreliable. This result indicates that these Asian banking sectors are operating in a very high efficiency level across the sample year. However, regarding to impact of the Asian financial crisis in 1997, it might be more reasonable to expect lower level of cost efficiency level since financial systems in countries like Indonesia, Malaysia, and Philippines, have suffered the destructive damage as economies shrank. Moreover, the reason for such a high efficiency score from 'true' SFA models can be explained by its methodological specification. As argued in the methodology section, although the introduction of 'true' SFA models can move some of the time-invariant heterogeneities out of inefficiency, they cannot distinguish these heterogeneities from time-invariant inefficiency. If there is persistent inefficiency, it will be completely absorbed in the firm-specific constant term as time-invariant heterogeneities, implying that only time-varying inefficiency is actually measured. This means that 'true' SFA models overcompensate for the heterogeneity and underestimate inefficiency, thereby overestimating cost efficiency. Unfortunately, this is the inherent feature of the modeling process since $\alpha_i + u_{it}$ contains both country-specific heterogeneities and inefficiency, and both may have time-invariant and time-varying elements. Furthermore, as suggested in Greene (2004), the 'true' random-effects model will be unstable if more variants are added into the model.

To summarize, classical panel data models assume all the time-invariant effects to be time-invariant inefficiency, which inevitably pushes all the time-invariant

[10] In Table 2, λ for true RE model is 382.997 but not significant at the 10% level, indicating the acceptance of null hypothesis that λ is 0, suggesting no inefficiency.

heterogeneities into inefficiency. Conversely, 'true' SFA models assume all the time-invariant effects to be time-invariant heterogeneities that also absorb time-invariant inefficiency. Therefore classical panel data models underestimate cost efficiency scores and 'true' SFA models overestimate them. However, this study intends to construct a panel data model that can include cross-country heterogeneities on purpose to examine whether cross-country differences are important to banks' performance. Moreover, it is also important to distinguish between time-invariant heterogeneities and time-invariant inefficiency to ensure no overcompensation for time-invariant heterogeneity and hence more accurate estimation of cost efficiency. Therefore, 'true' SFA models are not the appropriate model candidates. Furthermore, as the classical FE model uses LSDV in the estimation, it provides no means of including cross-country heterogeneities. Only the classical RE model can incorporate cross-country heterogeneities that are distinguished from time-invariant inefficiency. Because Battese and Coelli (1992) model allows inefficiency to vary over time, it is preferred to Pitt and Lee (1981) model and will be adopted in the next stage.

4.4.2 Cost efficiency estimates with incorporation of cross-country heterogeneities

Estimated efficiency results of Battese and Coelli (1992) model with incorporation of cross-country heterogeneities are reported in Table 4.5 with the estimated parameter coefficients presented in Table 4.6.

Table 4.5: Estimated average efficiency score from Battese and Coelli model with cross-country heterogeneities

	1998	1999	2000	2001	2002	2003	2004	2005	Average
Mean	0.6154	0.6226	0.6131	0.6000	0.5843	0.5778	0.5642	0.5561	0.5897
Std.dev.	0.1414	0.1705	0.1734	0.1766	0.1745	0.1786	0.1835	0.1896	0.1771
Minimum	0.3084	0.2238	0.2088	0.1941	0.1799	0.1661	0.1528	0.1400	0.1400
Maximum	0.9871	0.9905	0.9900	0.9896	0.9891	0.9886	0.9881	0.9875	0.9905
Sample Obs.	132	232	242	250	259	271	274	230	1890

The average cost efficiency score is about 0.5897 across the sample, an increase of 16 percentage points over the efficiency score estimated when cross-country heterogeneities are not included, suggesting that differences between countries are indeed important sources to explain banking performance in these Asian economies other than the managerial abilities of individual bank. If these cross-country heterogeneities do exist but are not included, their presences will be misclassified as inefficiency, leading to biased and unreliable efficiency estimates that could mislead policy makers. In addition, neutral technical progress is found in these Asian banking industries since the time coefficient is statistically significantly negative (-0.021, with a *t*-statistic of -4.349), which suggest that banks benefit from using new technology, such as the introduction of ATM, telephone banking and internet banking services. The overall cost efficiency in these Asian areas decreases from 0.6154 in 1998 to 0.5897 in 2005[11]. The overall decreasing trend of cost efficiency over the sample period is reflected by the value of η, estimated as -0.045 with a *t*-statistic of -14.958. As discussed in methodology section, in Battese and Coelli (1992) model, the statistically significant negative value of η suggests that the cost inefficiency score is increasing in an increasing rate. The decline in efficiency may not necessarily imply that banks' performance is deteriorating over time as the estimation does not account for the quality of outputs (e.g. quality of loans). As argued by Kumbhakar and Wang (2007), an improvement in quality of output (e.g. disposal of non-performing loans, NPLs) may appear as a reduction in efficiency. The Asian financial crisis left a large volume of NPLs for Asian banks and one important objective in the post-crisis financial and banking reforms was to reduce the volume of NPLs in bank assets. This may appear as a reduction in net loan values for a given level of inputs and, consequently may be reflected as a decrease in efficiency. Furthermore, the need to reduce the quantity of NPLs in the banking system may have forced governments to push for mergers between healthy and troubled banks, resulting in declined performance of existing healthy banks and a decline in efficiency scores. Table 4.7 presents cost efficiency scores across the countries and regions over the sample period. These show the same decreasing trend as the overall score, although we observe a few cases of efficiency improvement in some years, which can also partly be explained by the nature of our unbalanced panel data set.

[11] Although this observed decline in efficiency score is not continuous over the sample period. We observe a small increase from 0.6154 in 1998 to 0.6226 in 1999, caused by the smaller number of bank observations in 1998 than in 1999 due to the unbalanced sample data set.

Table 4.6: Estimated Battese and Coelli model with cross-country heterogeneities

Parameters	Coefficients	t-statistics
Constant	-0.171***	-4.675
Y1	0.540***	46.896
Y2	0.307***	22.205
Y3	0.048***	7.163
W1	0.716***	61.854
W2	0.154***	15.988
lnE	-0.051***	-6.565
Y11	0.061***	31.543
Y12	-0.141***	-23.981
Y13	0.014***	6.339
Y22	0.088***	23.097
Y23	-0.042***	-16.605
Y33	0.015***	13.036
W11	0.047***	9.558
W12	-0.077***	-9.147
W22	0.029***	6.995
Y1W1	-0.022***	-4.344
Y1W2	0.018***	3.335
Y2W1	0.050***	7.349
Y2W2	-0.035***	-5.095
Y3W1	-0.003	-0.978
Y3W2	-0.004	-1.448
T	-0.021***	-4.349
SQRT	0.001	1.603
Y1T	-0.003	-1.406
Y2T	-0.000	0.219
Y3T	0.004***	3.660
W1T	0.001	0.378
W2T	-0.002	-1.033
Z1	0.145***	13.108
Z2	-0.291***	-12.420
Z3	0.201***	3.738
Z4	-0.048***	-4.109
Z5	-0.088***	-2.776
Z6	-0.103***	-4.750
Z7	-0.076***	-5.012
Z8	-0.056**	-2.339
σ_u	0.756***	24.964
σ_v	0.102	
λ	7.420***	757.657
σ	0.763	
η	-0.045***	-14.958
L.L.F.	1028.156	

All variables are in log terms.
Notes: ***, ** denotes the rejection of the null at the 1% and 5% level

Table 4.7: Estimated cost efficiency for individual country and year

	1998	1999	2000	2001	2002	2003	2004	2005	Average
China	0.5795	0.5919	0.5904	0.5899	0.5748	0.5755	0.5656	0.5461	0.5860
Hong Kong SAR	0.6838	0.6506	0.6309	0.6186	0.6119	0.5991	0.5863	0.6032	0.6197
India	0.8307	0.7918	0.7843	0.7760	0.7656	0.7570	0.7495	0.7402	0.7661
Indonesia	0.6421	0.5726	0.567	0.5505	0.5335	0.5233	0.5120	0.4770	0.5376
Korea	0.5944	0.5741	0.5639	0.5459	0.5361	0.5210	0.4987	0.4760	0.5397
Malaysia	0.7303	0.6921	0.6805	0.6637	0.6351	0.6406	0.6287	0.6312	0.6666
Philippines	0.5879	0.5754	0.5609	0.5544	0.5377	0.5229	0.5080	0.5431	0.5475
Singapore	0.7550	0.7453	0.7352	0.7248	0.7140	0.7030	0.6916	0.7159	0.7224
Taiwan	0.5644	0.5517	0.5639	0.5174	0.5016	0.4919	0.4762	0.4635	0.5121
Thailand	0.3495	0.2927	0.2784	0.2618	0.2397	0.2179	0.2145	0.1999	0.2517
All	0.6154	0.6226	0.6131	0.6000	0.5843	0.5778	0.5642	0.5561	0.5897
Of Sample Obs.	132	232	242	250	259	271	274	230	1890

The average cost efficiency for each country and region suggests that India is the most efficient, with an efficiency score of 0.7661. Singapore, Malaysia and Hong Kong SAR follow with efficiency scores of 0.7224, 0.6666 and 0.6197 respectively. China, Korea, Indonesia and Taiwan rank fifth to ninth with efficiency scores of 0.5860, 0.5397, 0.5376 and 0.5121 respectively. Thailand is the least efficient with an efficiency score of 0.2517.

However, although it is difficult to compare this result with the existing literature due to the lack of other cross-country efficiency studies in Asia, comparisons based on individual country are doable. For instance, country-specific efficiency ranking indicates a tough challenge ahead of Chinese banks because of the ease of foreign entry and more open domestic financial and banking markets. Foreign commercial banks from other Asian countries and regions are likely to outperform the domestic Chinese banks with superior managerial ability, advanced technology and better risk management and monitoring skills, etc. Further banking reform is therefore needed, especially for the dominant state-owned commercial banks. Compared with joint-stock commercial banks, which are jointly owned by the state and private companies, the big five state-owned commercial banks (Industrial and Commercial Bank of China, China Construction Bank Corporation, Agricultural Banks of China, Bank of China Limited and Bank of Communication Co. Ltd), are about 25 percentage points less efficient on average from a cost perspective (see Figure 4.1). This finding is consistent with other Chinese banking efficiency studies. Berger *et al.* (2009) examine the profit and cost efficiency of Chinese banks from 1994 to 2003 and find that big four state-owned banks are the least efficient compared with non-state owned and foreign banks. In Kumbhakar and Wang's (2007) sample from 1993 to 2002, the authors find that whole state owned banks are about 43% less technically efficient than joint-stock banks using input distance function. Fu and Heffernan (2007) measure cost efficiency of 14 commercial banks from 1985 to 2002 and find that state owned banks were 9% less efficient than joint-stock banks. Although the magnitude of inefficiency remains vastly different, it is widely agreed that state ownership is the main reason of poor performance.

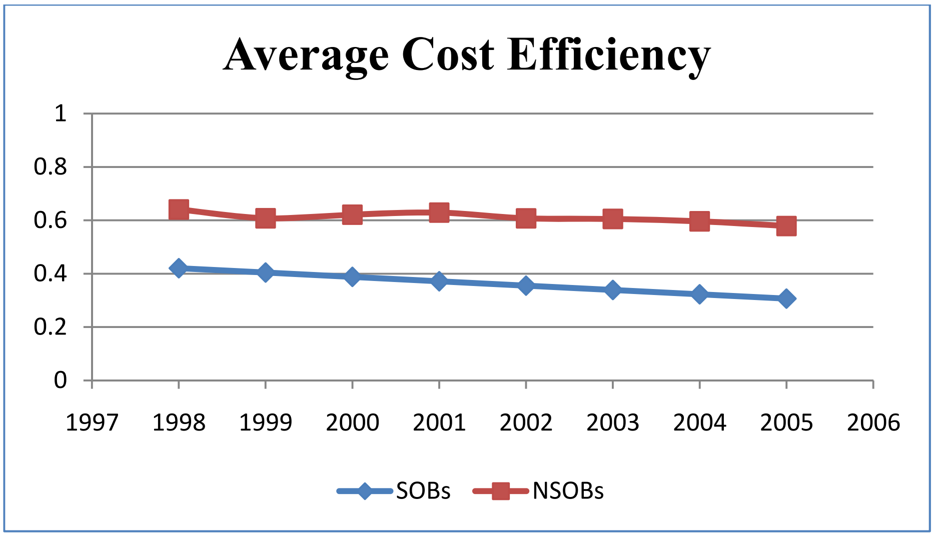

Figure 4.1: Trend of average cost efficiency of Chinese state-owned banks and non state-owned banks

Thus one set of policy recommendations is to accelerate deregulation and privatization of state-owned banks and to introduce comprehensive risk management and a supervisory system to prevent the excessive risk taking, and to mitigate principle-agent problems and moral hazard problems. Moreover, it is necessary to strengthen the independence of state-owned banks in decision making and credit allocation. Under the current political system, senior managers in state-owned banks are appointed by the government. Inevitably, to fulfill the requirements of local government, bank managers have to implement some 'policy-lending' that may depart from the shareholders' interest of maximizing profits and may result in NPLs, due to lack of credit assessment and monitoring.

It is also interesting to see that Chinese banks outperform Korean banks by a margin of 0.05. This could be explained in three aspects. First, this study focuses on retailing business rather than wholesale banking services. In the 1990s, in order to enhance competition, the Korean government eased the entry barriers for foreign banks. Due to their comparative advantage, these foreign banks operated quite successfully in wholesale business. However, since this study excludes such wholesale banking institutions, their success is not recognized in efficiency scores of Korean banks. Moreover, the Korean banking industry experienced substantial structural collapse during the 1997 financial crisis. In the post-crisis financial reconstruction, the Korean financial authority implemented a prompt corrective action (PCA) system in order to reduce NPLs, to enhance Basel capital adequacy ratios and to strengthen prudential regulation. Through the injection of public funds for purchase of NPLs and recapitalization of banks, NPLs ratio declined from 13.6% at the end of 1999 to 2.0% at the end of 2004 (Kim *et al.*, 2006). Banks with a large volume of NPLs were forced to exit or to merge with healthy banks, leaving only 19 banks in 2005 compared to 33 banks at the end of 1997. Therefore, the profitability and performance of healthy banks is likely to be weakened during the post-crisis period, reflected as a decline in the efficiency score. The most important part of the PCA system is to ensure Korean banks to satisfy the Basel capital adequacy ratio of 8%. The capital adequacy ratio increased from 7% by the end of 1997 to 11.3% at the end of 2004. Although a strong capital adequacy ratio can reduce insolvency risks, it induces a high cost for Korean banks. In contrast, the capital ratio for Chinese banks is only 4% over the sample period since China announces that it would not adopt the Basle Accord II of 8% until 2008. The higher adequacy ratio for Korean banks may have reduced the efficiency level from a cost perspective.

As discussed in section 4.2, properties of the translog cost function need to be satisfied to ensure that cost efficiency estimates are reliable. These results are reported in Table 4.8, with all the required properties checked at the sample mean and across the whole sample. Monotonicity conditions are strongly satisfied at the sample mean, where the elasticities of output and input prices are statistically significantly higher than zero, suggesting that the cost function is non-decreasing in $\mathbf{y}$ and $\mathbf{w}$. In addition, they are satisfied at the majority of the sample points. Concavity condition is satisfied at the sample mean and across 95.5 percent of the sample points, since the Hessian matrix with respect to the input prices should be negative semi-definite. The satisfaction of monotonicity and concavity properties indicates that the fitted model is a true cost function and that cost efficiency estimates are reliable. The scale elasticity at the sample mean is 0.895 and is statistically significantly different from one, suggesting slight scale economies in these Asian banking industries. This finding coincides with the evidence of scale economies in the US banking industries when accounting for risk and financial capital. Berger and Mester (1997) investigate about 6000 commercial banks over 1990-1995 and estimate scale economies and X-efficiency for banks of different sizes, finding significant cost scale economies in each size class in their preferred model. Hughes and Mester (1998) also find scale economies across all bank sizes in a sample from 1989 to 1990.

Table 4.8: Properties (monotonicity, scale and concavity) of the fitted cost function adopting Battese and Coelli random effects with cross-country heterogeneities at the sample mean and throughout the sample

Monotonicity Property	**Elasticity**	**Parameters**	**Standard errors**	**Whole sample: % of sample points with cost increasing in output and input prices**
at the sample mean	ey1	0.540	0.036	99.6
at the sample mean	ey2	0.307	0.012	99.6
at the sample mean	ey3	0.048	0.014	92.7
at the sample mean	ew1	0.716	0.007	100
at the sample mean	ew2	0.154	0.012	99.6
Scale Property	**Scale Elasticity**	**Standard errors**	**t-value**	**Whole sample: % of sample points with increasing returns to scale**
at the sample mean	E = 0.895	0.013	-8.320 Reject H_0: E=1	99.2
Concavity Property	**Objective function**	**Principle Minors**	**Values**	**Whole sample: % of sample where H(w) is negative definite**
at the sample mean	H(w)	First order	-0.110, -0.071	95.5
		Second order	0.007	

4.4.3 Impact of cross-country heterogeneous factors

Including cross-country environmental variables help correctly construct the cost frontier and exclude some part of time invariant heterogeneities out of inefficiency. Besides these impacts, it is also interested to check whether the influences of these environmental variables are in line of the expectation. This is reported in Table 4.9.

Table 4.9: The expected and observed influences of environmental variables on banks' costs

	Main Economic Indicators				**Banking Structure**			
	PD	GP	IF	UN	BC	NI	CR	IR
Expected	**-**	+ **or -**	+	**-**	+ **or -**	**-**	**-**	**-**
Battese and Coelli	+	**-**	+	**-**	**-**	**-**	**-**	**-**

Notes: PD= Population density; GP= GDP per capita; IF= Inflation; UN= Unemployment; BC= Banking concentration; NI= Net interest margin; CR= Capital ratio; IR= Intermediation ratio

First, consider the role of main economic indicators. Contrary to the expectation, the coefficient of population density variables has a positive sign. Higher density contributes to an increase in banking costs, instead of the expected decrease in costs. One reason can be found in the characteristics of banking competition. In higher density area, banks may be forced to open more branches to compete for customers. Other promotion and strategic operations may also increase the level of banks' cost. The expectation for GDP per capita can be either positive or negative. Coinciding with this expectation, negative sign is observed, suggesting that banks benefit more from the technological change and well diversification and expansion of their business, which substantially reduce their operational expenses. The positive sign of inflation is also in line with the expectation as the higher inflation, the higher costs it may incur since the inflation may increase input prices involved in the banking production process. For instance, employees may demand higher payment and savers may ask for higher deposits rate, etc. Last candidate in this group is unemployment ratio, the other main indicator for the macro-economic environment. High unemployment rate reflects poor economic development. Therefore, banks in these countries may seek various routes to maintain the current business level and reduce the potential risks in their loans and other services. At the meanwhile, it may also cut the costs sharply to balance the income sheet. Therefore, the expectation of the influence of unemployment is negative, which is also reflected in empirical results.

The second group of environmental variables is those reflect the banking structure and regulatory conditions. The first is banking concentration. As discussed before, higher concentration may be associated with either higher or lower costs. On the one hand, if higher concentration is a result of market power, one may expect costs go in the same direction (Leibenstein, 1966). The author points out that some of the empirical evidences suggest that it is possible that the lack of competitive pressure of operating in monopoly industries would lead to a higher cost than would be the case under competition. It is also highly likely that firm managers are capable of reduce the cost if the environment forced them to do so. However, on the other hand, if higher concentration results from the superior management, we may expect a negative sign, suggesting higher concentration can be associated with lower costs (Demsetz, 1973). As argued by Demsetz, under the environment of competition, where absence of effective barriers to entry is apparent, an industry could be concentrated only if some firms are superior in producing and marketing products, which enable them to have a differential advantage in expanding output develops in the firms. Such expansion will increase the concentration of the industry, as well as the increases in the rate of return and cost advantage these firms earn. Such cost advantage may be reflected in economics of scale or in downward shifts in positively sloped marginal cost curves. The empirical results show the negative sign in favour of the superior management. The second indicator is net interest margin, measured as the difference between interest rates associated with loans and deposits. The higher interest margin, the more profitable is the bank and higher ability to convert deposits to loans, the lower are banks' costs, as is observed in the empirical results. The negative sign of capital ratio is also expected since lower capital ratio indicates higher insolvency risk of banks therefore banks may have a higher operational cost in running the business. The last environmental variable is the intermediation ratio. It captures the ability of banks to convert the deposits into loans. The higher intermediation ratio indicates the higher ability of repay the interests to depositors and then may lower banks' costs. As expected, a negative sign is observed.

4.5 Discussions

4.5.1 Choice of output and input prices specifications

As surveyed in chapter 2, different approaches of output and input specifications have been used in efficiency studies. Advantages and disadvantages of these approaches have been pointed out clearly but still there is no consensus which approach is the most appropriate and the choice is made upon what issue and data are examined. However, it is widely agreed that neither intermediation approach nor production approach fully captures the dual role of deposits. As suggested by Berger and Humphrey (1997), a dual approach may be the better candidate due to its acknowledgement of dual role of deposits by considering deposits as both output and input. However, in the existing literature, although the dual approach has been applied in different banking efficiency studies, its credibility has never been tested, especially when stochastic frontier approach is adopted. The main focus of those studies is on obtaining efficiency scores without checking the statistical inferences and theoretical properties associated with the adopted functional form. Table 4.10 provides a robustness check list for those studies using dual approach. Only two of them present the parameter estimates and either p-value or t-statistics for robustness check. Moreover, none of them check the theoretical properties for using cost function or profit function, as what is done in this work. The satisfaction of these properties ensures the credibility of using either cost function or profit function to estimate cost or profit efficiency. Without this robustness test, the estimated efficiency level might be misleading and so as the correspondent policy implications and suggestions, which may ultimately result in a wrong direction of administrative and operational adjustments.

Table 4.10: Studies using dual approaches

Studies	SFA	Parameter estimates provided	Statistical inferences provided	Properties of functional form checked
Berger *et al.* (2009)	Yes	No	No	No
Carvallo and Kasman (2005)	Yes	No	No	No
Fu and Heffernan (2007)	Yes	Yes	Yes	No
Hao *et al.* (2001)	Yes	Yes	Yes	No
Hasan and Marton (2003)	Yes	No	No	No

Before providing these test results from my sample data, it is worth checking whether the existing two studies using dual approach can pass the test. Apparently, Fu and Heffernan (2007) failed the robustness test in three aspects. Key parameter estimates from their paper that are essential to run the robust test are provided in Table 4.11[12].

Table 4.11: Key parameter estimates from Fu and Heffernan (2007)

Parameters	Half-Normal		Exponential		Truncated Normal	
	Coefficients	p-value	Coefficients	p-value	Coefficients	p-value
Outputs						
Total Deposits	0.489	0.3439	0.550	0.2780	0.582	0.2632
Total Loans	0.577	0.3375	0.649	0.2809	0.778	0.2042
Total Investments	0.149	0.0369	0.152	0.0342	0.163	0.0274
Non-Interest Income	0.109	0.2914	0.102	0.3099	0.103	0.3183
Input Prices						
Price of funds	1.152	0.0000	1.170	0.0000	1.193	0.0000
Price of fixed assets	-0.447	0.0974	-0.465	0.0759	-0.493	0.0644
2^{nd} order derivatives with respect to input prices						
Price of funds	0.131	0.0000	0.134	0.0000	0.136	0.0000
Interactive term	-0.087	0.0467	-0.092	0.0324	-0.094	0.0352
Price of fixed assets	0.082	0.2612	0.085	0.2311	0.088	0.2240

All variables are in log terms.
Source: Fu and Heffernan (2007)

First of all, for all three distributional assumptions, p-values for three out of four outputs (total loans, total deposits, non-interest income) are not significant at the 10% level. Second, the theory of monotonicity condition of cost function requires that total costs to be non-decreasing in input prices, indicating that the first order derivatives of total costs with respect to input prices must be positive. However, coefficient of the first order derivative with respect to price of fixed assets is negative, indicating the failure of satisfaction of monotonicity conditions. Finally, regarding to concavity condition of cost function, total costs are concave in input prices. This condition is satisfied when Hessian matrix of input prices are negative semi-definite. Using their second order derivative with respect to input prices, we can construct the Hessian matrix and check whether concavity condition is satisfied at the sample mean. The results are presented in

[12] For the full list of parameter estimates, please refer to their paper.

Table 4.12, which clearly indicates the failure of concavity condition because for Hessian matrix to be negative semi-definite, since first order principle minors must be negative and second order principle minors must be positive. Key parameter estimates from Hao *et al.* (2001) are provided in Table 4.13. Only demanding deposits and wage rate are significant at the 5% and 10% level, respectively. It fails the monotonicity and concavity test as well because first order derivatives of total costs with respect to demanding deposits and price for physical capital are negative and Hessian matrix is not negative semi-definite (see Table 4.14).

Table 4.12: Hessian matrix for Fu and Heffernan (2007)

Principle Minors	Half-Normal	Exponential	Truncated Normal
First order	0.4371	0.4669	0.5022
	0.8108	0.8512	0.9120
Second order	-0.0079	-0.0071	-0.0073

Table 4.13: Key parameter estimates from Hao *et al.* (2001)

Parameters	Coefficients	Parameters	Coefficients
Outputs		Input Prices	
Total Loans and Securities (Y1)	0.575	Wage rate (W1)	16360*
Demanding Deposits (Y2)	-0.188**	Interest for borrowed funds (W2)	71462
Fee Income (Y3)	1.570	Price of physical capital (W3)	-7045.9
		2^{nd} derivatives with respect to input prices	
		W1*W1	-355.41**
		W2*W2	42769
		W3*W3	25337
		W1*W2	-24561
		W1*W3	-14765
		W2*W3	-15774

All variables are in log terms. **, * denotes the rejection of the null at the 5% and 10% level.
Source: Hao *et al.* (2001)

Table 4.14: Hessian matrix for Hao *et al.* (2001)

	Principle Minors
First order	0.267632E+09
	0.510679E+10
	0.496771E+08
Second order	-8613.370

Table 4.15: Output and input prices specification

	Production approach	Intermediation approach	Dual approach
Outputs	Total loans (Y1) Other earning assets (Y2) Total deposits (Y3) Non-interest income (Y4)	Total loans (Y1) Other earning assets (Y2) Non-interest income (Y3)	Total loans (Y1) Other earning assets (Y2) Total deposits (Y3) Non-interest income (Y4)
Input prices	Price of labour (W1) Price of fixed assets (W2)	Price of funds (W1) Price of labour (W2) Price of fixed assets (W3)	Price of funds (W1) Price of labour (W2) Price of fixed assets (W3)

Table 4.16 presents empirical estimates using production approach and dual approach to measure cost efficiency of ten Asian banking industries from 1998 to 2005. The main output and input prices specifications are provided in Table 4.15. Therefore, for production approach, four outputs and two input prices variables are specified, while for dual approach, four outputs and three input prices variables are defined. The cost efficiency estimate from production approach is 0.3415 on average, which is about 25 percentage points less than the ones from intermediation approach. This can be partly explained by the poor explanatory power of the cost regression as two main outputs are insignificantly negative. Moreover, the negativity of coefficients for total loans and other earning assets also fails the monotonicity test both at the sample mean and over the entire sample (see Table 4.17). Despite the satisfaction of monotonicity condition at the sample mean, it is not guaranteed over the entire sample for non-interest income and total deposits variables.

The overall poor empirical results from using production approach is expected since we are actually using value of total deposits rather than the number of deposit accounts suggested in the literature. This explains the domination of the coefficient of total deposits in the output derivatives. Ferrier and Lovell (1990) is the only study using production approach and the authors use the number of demand deposits account and the number of time deposits account as measures of total deposits. In their empirical results (Ferrier and Lovell, 1990: 240-241), although the coefficient for commercial loans is not statistically significant, all outputs and input prices variables satisfy the monotonicity condition at the sample mean, despite the fact of dissatisfaction of concavity condition at the sample mean (see Table 4.18). This indicates that when using

Table 4.16: Parameter estimates for production approach and dual approach

	Production approach		Dual approach	
Parameters	**Coeff.**	**t-ratio**	**Coeff.**	**t-ratio**
Constant	-0.7452***	-16.057	-0.1096***	-3.438
Y1	-0.0581	-1.299	0.0881***	5.224
Y2	-0.0292	-0.611	-0.0041	-0.233
Y3	0.0561***	4.206	0.039***	7.151
Y4	0.9309***	10.696	0.8111***	28.645
W1			0.7148***	78.268
W2	0.6698***	51.136	0.1634***	19.563
lnE	0.0388*	1.804	0.0353***	4.837
Y11	0.0511***	4.448	0.0204***	4.161
Y12	0.0423	1.098	0.0403**	2.337
Y13	0.0263**	2.077	0.0207***	3.215
Y14	-0.2284***	-3.954	-0.0951***	-5.189
Y22	0.0034	0.172	0.0245**	2.203
Y23	-0.016	-0.878	0.0111*	-1.932
Y24	-0.0456	-0.660	-0.0869***	-2.960
Y33	0.016***	5.642	0.0128***	13.086
Y34	-0.0374	-1.427	-0.036***	-3.455
Y44	0.1766***	2.639	0.107***	4.843
W11			0.0446***	10.537
W12			-0.0663***	-8.690
W22	0.001***	2.362	0.0238***	6.607
Y1W1			-0.0395***	-4.470
Y1W2	0.0152	1.292	0.0277***	2.674
Y2W1			-0.0006	-0.062
Y2W2	0.0065	0.477	0.008	0.671
Y3W1			-0.0058**	-1.976
Y3W2	-0.0079*	-1.820	-0.0002	-0.095
Y4W1			0.0691***	4.183
Y4W2	-0.0284	-1.390	-0.0494**	-2.531
T	-0.0815***	-8.627	-0.032***	-7.722
SQRT	0.0031**	2.409	0.0006	0.839
Y1T	-0.0067	-0.889	-0.0212***	-5.147
Y2T	-0.0159**	-2.483	-0.0112***	-3.069
Y3T	0.0029	1.527	0.0043***	4.812
Y4T	0.0204*	1.692	0.028***	4.422
W1T			-0.0001	-0.039
W2T	-0.0092***	-4.134	-0.0009	-0.434
Z1	0.2809***	10.368	0.1279***	13.484
Z2	-0.6036***	-10.724	-0.2842***	-16.604
Z3	1.2403***	11.719	0.1081*	1.920
Z4	-0.1267***	-4.675	0.001	0.093
Z5	-0.002	-0.033	-0.0491**	-2.226
Z6	0.1539***	4.225	-0.0742***	-3.508
Z7	-0.1048***	-5.120	-0.0418***	-3.227
Z8	0.0758*	1.606	-0.0016	-0.065

Table 4.16: Model estimates for production approach and dual approach (continued)

	Production approach		Dual approach	
Parameters	**Coeff.**	**t-ratio**	**Coeff.**	**t-ratio**
λ	6.2701***	439.115	9.2804***	1019.457
σ_u	1.218***	7.369	0.7438***	22.304
η	0.037***	9.988	-0.068***	-17.665
Cost efficiency				
Mean	0.3415		0.6052	
Std. Dev.	0.1950		0.1615	
Minimum	0.0468		0.1173	
Maximum	0.9803		0.9935	

Notes: ***, **, * denote the rejection of the null at the 1%, 5% and 10% level.

Table 4.17: Properties of cost functions using production and dual approach

		Production approach		Dual approach	
		Coeff.	**Satisfied**	**Coeff.**	**Satisfied**
Sample mean					
Monotonicity (1st order derivatives)	ey1	-0.0581	No	0.0881	Yes
	ey2	-0.0292	No	-0.0041	No
	ey3	0.0561	Yes	0.0390	Yes
	ey4	0.9309	Yes	0.8111	No
	ew1			0.7148	Yes
	ew2	0.6698	Yes	0.1634	Yes
Concavity Hessian matrix H(w)	1st order principle minors			-0.1146 -0.0891	Yes
	2nd order principle minors			-0.0158	No
Whole sample		**Sample %**		**Sample %**	
Monotonicity	ey1	11.7		28.8	
	ey2	11.7		17.5	
	ey3	27.7		22.0	
	ey4	71.3		84.2	
	ew1			100	
	ew2	100		91.4	
Concavity	Hessian H(w)			0.01	

production approach and considering deposits as banks' output, the number of total deposits accounts should be used. Therefore when data on the number of total deposits account is not available, using the production approach may not be the right choice and considering deposits as banks' output in the estimation may generate misleading results. For the same reason, researchers should also be concerned about the use of the dual approach that considers the dual role of deposits in the financial institutions as both output and input. As deposits account plays the role in providing transactions or document processing services for financial institutions, the number of deposits account should be considered as one of banks' output rather than the value of deposits used in all the studies adopted dual approach. This misunderstanding of the role of deposits as an output will not be a problem during the estimation of efficiency unless certain theoretical and statistical properties have been checked. However, unfortunately, those studies using dual approaches fail the robustness test as discussed before and the similar results can be found using my sample, in which the value of total deposits is considered as one of the four outputs. Although cost efficiency estimate is similar to the one from using intermediation approach in terms of mean, standard deviation, minimum and maximum of the efficiency level, the statistical and theoretical test in using dual approach provides distinctive results. Besides the coefficient of other earning asset is not significant, monotonicity and concavity condition are not satisfied both at the sample mean and over the sample.

To summarize, this sub-section discusses the choice of using different approaches to specify output and input variables. Based on comparisons of empirical results and robust test among production approach, intermediation approach and dual approach, it can be concluded that although it is ideal to consider the dual role of deposits as both output and input, without the availability of data on the number of total deposits accounts, using dual approach may not be the appropriate choice as it may create misleading results on efficiency estimates and wrong policy implication. If only data on values of total deposits is available as seen in most of banking data, intermediation approach may be the appropriate choice in regard to the satisfaction of statistical and theoretical properties.

Table 4.18: Monotonicity and concavity condition at the sample mean for Ferrier and Lovell (1990)

Parameters	Coefficients	Monotonicity
Outputs		
Number of demand deposit accounts (Y1)	0.316***	Yes
Number of time deposit accounts (Y2)	0.577***	Yes
Number of real estate loans (Y3)	0.029***	Yes
Number of installment loans (Y4)	0.052***	Yes
Number of commercial loans (Y5)	0.007	Yes
Input Prices		
Price of labour (W1/W3)	0.309***	Yes
Occupancy costs and expenditure on furniture and equipment (W2/W3)	0.213***	Yes
Expenditure on materials (W3)		
2nd derivatives with respect to input prices		
W1*W1	0.125***	
W1*W2	-0.035***	
W2*W2	0.076***	
Concavity	**Principle Minors**	**Satisfaction**
1st order	0.0364	No
	-0.0156	
2nd order	-0.0015	No

All variables are in log terms. *** denotes the rejection of the null at the 1% level.
Source: Ferrier and Lovell (1990)

4.5.2 Choice of functional forms: flexibility or credibility?

The theoretical argument for and against the use of different functional form has been provided in chapter 2. In this study, the translog cost function is adopted to measure the cost efficiency. First of all, it is more flexible than the Cobb-Douglas cost functional form as argued by Coelli *et al.* (2005) and Kaparakis *et al.* (1994) that it imposes few restrictions on the first- and second-order effects and at the same time, it can also be viewed as a second-order logarithmic approximation to an arbitrary continuous transformation surfaces. Second, the use of translog cost functional form dominates the existing literature. Measuring cost efficiency using translog functional form can help compare my empirical results with the existing efficiency studies. The dominance of translog functional form to Cobb-Douglas is also proved in my sample. Based on the likelihood ratio (LR) test, translog cost function is preferred than Cobb-Douglas cost function as the likelihood ratio is 862.937 (the critical value for chi-square distribution with 21 degrees of freedom is 38.93 at the 1% level and 32.67 at the 5% level).

However, McAllister and McManus (1993) question the suitability of translog functional form by arguing that fitting translog cost function, which was originally developed as a local approximation of some unknown true cost function, globally (i.e. across the large variety of bank sizes and output mix), will systematically misrepresent the costs of certain banks, resulting in a specification bias that contributes to the false conclusion of economies of scale, that may mislead policy makers who rely on those conclusion. Instead, the authors suggest alternative attempts to fit the cost function globally, for instance, using Kernel regression techniques, or the linear spine estimation techniques, or the Fourier approximation approaches. Since then, a number of studies attempt to achieve the globally approximation of unknown cost function by adopting the Fourier flexible (FF) functional form to measure efficiency and economies scale and scope, in which a series of linear combination of sine and cosine functions, so called Fourier series, are added into translog functional terms. Since the sine and cosine functions are mutually orthogonal, the Fourier series will represent the global approximation to an unknown true function. Therefore, it would be an interesting and useful addition if a discussion on the use of the FF form can be included in this Asian banking cost efficiency study.

4.5.2.1 The FF functional form and its application in banking efficiency studies

The FF functional form, first proposed by Gallant (1981), is believed to be able to approximate any function well over the entire range of the data, so called global approximation. The widely used popular translog functional form, representing the second-order Taylor series expansion, only provides the local approximation, and hence, biased estimates will be generated. The technical demonstration of the FF functional form can be seen in Gallant (1981, 1982) and Gallant and Souza (1991).

It is not until recently has the FF functional form been heavily adopted in the banking efficiency studies due to the nature of widely varied banking data in terms of bank size and output mix. The earliest attempt to use the FF functional form is McAllister and McManus (1993), in which the authors find that previous findings of economies of scale in small banks with total assets up to $100 million and decreasing returns of scale for large banks is inaccurate due to misspecification of using translog functional form. They argued that by using the FF terms, economies of scale are found up to $500 million in total assets and constant returns to scale are associated with larger banks. Another evidence of misspecification of translog functional form is found in Mitchell and Onvural (1996), in which no gains or small gains of economies of scale and scope for large commercial banks are found by estimating the FF cost function.

However, other studies attempting to use the FF functional form provide different results. Berger and Mester (1997) use different functional forms and conclude that no large differences between translog functional form and the FF functional form, although the coefficients on the Fourier terms are jointly significant at the 1% level. The relevant efficiency estimates and scale economies are generally the same. Altunbaş *et al.* (2001) find that average inefficiencies in European banking appear to range between 20% and 25% across the different size classes. This finding is consistent with Molynuex, *et al.* (1996) and Vander Vennet (2002), which indicates the little difference in using both the translog and Fourier flexible functional form. Carvallo and Kasman (2005) intend to use the Fourier flexible functional form but they fail to reject the null hypothesis of all Fourier parameters jointly equal to zero. Therefore, they use translog specification of the cost function. Fu and Heffernan (2007) argued that the Fourier flexible specification requires more degrees of freedom but with only a few banks and a short history, the Chinese data are limited by comparatively few observations. Therefore, they adopt translog functional form rather than the Fourier flexible functional form.

4.5.2.2 Existing approaches and methodological specifications

The FF functional form can be written as a nested function of translog function of outputs **y** and input prices **w**, and a vector of environmental variables **z**, plus a Fourier series of linear combination of sine and cosine functions of **x**, which is the vector of rescaled values of output y and input prices w. k_h determines the vector of Fourier series in the function and ε_{it} is the composed error terms.

$$\ln C = TL(\mathbf{y}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}) + \sum_{h=1}^{H} \left[\phi_h \cos(\mathbf{k}_h^{'} \mathbf{x}) + \varphi_h \sin(\mathbf{k}_h^{'} \mathbf{x}) \right] + \varepsilon$$

[4.11]

Across the literature of banking efficiency studies adopting the FF functional form, three approaches can be recognized in ways of constructing the Fourier series. These three approaches differ from each other in how **x** is determined.

The first approach is applied in Mitchell and Onvural (1996) (MO approach), in which the author using the following the specifications:

$$\ln C = \alpha_0 + \mathbf{b'x} + \frac{1}{2}\mathbf{x'Ax} + \sum_{h=1}^{H} \left[u_h \cos(\mathbf{k}_h^{'} \mathbf{x}) + v_h \sin(\mathbf{k}_h^{'} \mathbf{x}) \right] + \varepsilon$$

[4.12]

where **x=(l, z)** is an J+M vector of scaled log input prices output quantities, defined as $\mathbf{l} = \lambda \cdot (\ln \mathbf{w}_{it} + \mathbf{w}_w), z = \lambda \cdot \mu \cdot (\ln \mathbf{y}_{it} + \mathbf{w}_y)$, where $(\lambda, \mu, \mathbf{w}_w, \mathbf{w}_y)$ are scale factors**.** Therefore, the Fourier series k_h will include:

$\cos(z_j), \sin(z_j), \cos(z_j + z_r), \sin(z_j + z_r), \cos(l_m\text{-}l_n), \sin(l_m\text{-}l_n), \cos(l_m\text{-}l_n + z_j), \sin(l_m\text{-}l_n + z_j)$

The use of the FF form requires that the data be scaled so the difference between the maximum and minimum values of each independent variable does not exceed 2π . The scaling technique is the one suggested by Gallant (1982), described as procedure 1.

Procedure 1

$\mathrm{w}_{\mathrm{m}}^{\min}$ = sample minimum value of the mth input prices, m = 1, 2, 3
$\mathrm{w}_{\mathrm{m}}^{\max}$ = sample maximum value of the mth input prices, m = 1, 2, 3
$\mathrm{y}_{\mathrm{j}}^{\min}$ = sample minimum value of the jth output, j = 1, 2, 3
$\mathrm{y}_{\mathrm{j}}^{\max}$ = sample maximum value of the jth output, j = 1, 2, 3
$\theta_{\mathrm{pm}} = 0.00001 - \ln \mathrm{p}_{\mathrm{m}}^{\min}$
$\theta_{\mathrm{yj}} = 0.00001 - \ln \mathrm{y}_{\mathrm{j}}^{\min}$
M = sample maximum value of $\ln \mathrm{w}_{\mathrm{m}}^{\max} + \theta_{\mathrm{wm}}$, m = 1, 2, 3
$\lambda = 6/\mathrm{M}$
$\mu_{\mathrm{j}} = 6/\left[\ln \mathrm{y}_{\mathrm{j}}^{\max} + \theta_{\mathrm{yj}}\right] \cdot \lambda$, j = 1, 2, 3
scaled $\ln \mathrm{w}_{\mathrm{m}} = \lambda\left[\ln \mathrm{w}_{\mathrm{m}} + \theta_{\mathrm{wm}}\right]$
scaled $\ln \mathrm{y}_{\mathrm{j}} = \lambda\mu'\left(\ln \mathrm{y}_{\mathrm{j}} + \theta_{\mathrm{yj}}\right)$

The second approach is adopted by Berger and Mester (1997) (BM approach),

$$\ln C = \alpha_0 + \boldsymbol{\beta}'\mathbf{y} + \boldsymbol{\gamma}'\mathbf{w} + \frac{1}{2}\mathbf{y}'\mathbf{A}\mathbf{y} + \frac{1}{2}\mathbf{w}'\mathbf{B}\mathbf{w} + \frac{1}{2}\mathbf{y}'\mathbf{C}\mathbf{w} + \sum_{h=1}^{H}\left[\phi_h \cos\left(\mathbf{k}_h^{'}\mathbf{x}\right) + \varphi_h \sin\left(\mathbf{k}_h^{'}\mathbf{x}\right)\right] + \varepsilon \qquad [4.13]$$

where $\mathbf{x} = (\mathbf{y}, \mathbf{w})'$ is a J+M vector of scaled output and input prices, which guarantee $\mathbf{x} \sim [0, 2\pi]$. The Fourier series k_h will include:

$\cos(x_{yj}), \sin(x_{yj}), \cos(x_{yj} + x_{yr}), \sin(x_{yj} + x_{yr}), \cos(x_{wm}), \sin(x_{wm}), \cos(x_{wm} + x_{wn}),$
$\sin(x_{wm} + x_{wn}), \cos(x_{yj} + x_{wm}), \sin(x_{yj} + x_{wm})$

In their attempts, the scaling procedure of x can be written as procedure 2.

Procedure 2

Cutting 10% from each end of $[0, 2\pi]$ interval so that x_{j+m} will span $[0.1 \times 2\pi, 0.9 \times 2\pi]$ to reduce the approximation problems near the endpoints.
The formula for x_{j+m} is $0.2\pi - \mu \times a + \mu \times$ variable, where $[a, b]$ is the range of variable being tranformed, and $\mu = (0.9 \times 2\pi - 0.1 \times 2\pi)/(b - a)$.

The third attempt is used by Altunbaş *et al.* (2001) (AGMM approach), in which the authors combine the above two approaches.

$$\ln C = \alpha_0 + \boldsymbol{\beta}'\mathbf{y} + \boldsymbol{\gamma}'\mathbf{w} + \frac{1}{2}\mathbf{y}'\mathbf{A}\mathbf{y} + \frac{1}{2}\mathbf{w}'\mathbf{B}\mathbf{w} + \frac{1}{2}\mathbf{y}'\mathbf{C}\mathbf{w} + \sum_{h=1}^{H}\left[\phi_h \cos\left(\mathbf{k}_h^{'}x\right) + \varphi_h \sin\left(\mathbf{k}_h^{'}x\right)\right] + \varepsilon \qquad [4.14]$$

where $\mathbf{x} = (\mathbf{y}, \mathbf{w})'$ is also a J+M vector of scaled output and input prices, which guarantee $\mathbf{x} \sim [0, 2\pi]$. The Fourier series k_h will include:

$\cos(x_{yj}), \sin(x_{yj}), \cos(x_{yj} + x_{yr}), \sin(x_{yj} + x_{yr})$

This is the same model specification as used in BM approach but however, they differ in the scaling procedure. AGMM approach use the same scaling procedure as in MO approach. Be aware, the independent variable **y** and **w** will be scaled before estimation. In MO approach, **y** and **w** will be scaled using procedure 1, the same as the Fourier terms. However, in BM and AGMM approach, the independent variable **y** and **w** will be scaled using mean correction technique but the Fourier terms will be scaled using procedure 2 and procedure 1 respectively. As usual, all the theoretical properties associated with cost function need to be satisfied (i.e. monotonicity, concavity and homogeneity of degree one).

4.5.2.3 Empirical comparisons

In this section, the above three FF approaches are applied to the Asian banking data on purpose to examine whether FF functional form outperform the translog functional form. As discussed in last section, data has to be scaled prior to the estimation. Data transformation of MO approach is summarized in Table 4.19, while Table 4.20 and 4.21 describes the transforming process for BM approach and AGMM approach.

Table 4.22 and 4.23 present the cost efficiency results for different panel data models estimated from adopting MO approach, BM approach and AGMM approach. The results suggest the similar empirical findings when the translog functional form is adopted. Without including cross country heterogeneities, the classical panel data models will underestimate the cost efficiency scores, while 'true' SFA models will overestimate them due to the overcompensation for the heterogeneities. Including cross-country heterogeneities provides the improved cost efficiency scores, suggesting that cross-country differences are important and should not be neglected.

Table 4.19: Scaled variables for the Fourier flexible functions: MO approach

	Sample		Sample
θ_{w1}	2.6181	λ	1.6728
θ_{w2}	2.7065	μ_1	0.6168
θ_{w3}	2.8258	μ_2	0.6714
θ_{y1}	-0.4183	μ_3	0.7088
θ_{y2}	-0.7912		
θ_{y3}	0.8718		

	Sample mean	Minimum	Maximum
scaled $\mathrm{w}_1 = (\ln w_1 + \theta_{w1}) \cdot \lambda$	2.0140	0.00002	3.79523
scaled $\mathrm{w}_2 = (\ln w_2 + \theta_{w2}) \cdot \lambda$	2.1163	0.00002	4.08848
scaled $\mathrm{w}_3 = (\ln w_3 + \theta_{w3}) \cdot \lambda$	4.2256	0.00002	6.00000
scaled $\mathrm{y}_1 = (\ln y_1 + \theta_{y1}) \cdot \mu_1 \cdot \lambda$	3.4151	0.00001	6.00000
scaled $\mathrm{y}_2 = (\ln y_2 + \theta_{y2}) \cdot \mu_2 \cdot \lambda$	3.1279	0.00001	6.00000
scaled $\mathrm{y}_3 = (\ln y_3 + \theta_{y3}) \cdot \mu_3 \cdot \lambda$	3.2968	0.00001	6.00000

Table 4.20: Scaled variables for the Fourier flexible functions: BM approach

	Sample		Sample
μ_{w1}	0.462039	μ_{y1}	0.37541
μ_{w2}	0.502661	μ_{y2}	0.408612
		μ_{y3}	0.431371
	Sample mean	**Minimum**	**Maximum**
scaled w_1	2.5304	0.62832	5.65487
scaled w_2	2.8342	0.62832	5.65487
scaled y_1	3.4893	0.62832	5.65487
scaled y_2	3.2487	0.62832	5.65487
scaled y_3	3.3929	0.62832	5.65487

Table 4.21: Scaled variables for the Fourier flexible functions: AGMM approach

	Sample		Sample
θ_{w1}	3.88737	λ	0.55147
θ_{w2}	4.14551	μ_{y1}	0.81252
θ_{y1}	9.41966	μ_{y2}	0.88438
θ_{y2}	8.14111	μ_{y3}	0.93363
θ_{y3}	7.81578		
	Sample mean	**Minimum**	**Maximum**
scaled w_1	3.4153	0.00045	6.00000
scaled w_2	3.1281	0.00045	6.00000
scaled y_1	3.3002	0.00045	6.00000
scaled y_2	2.2707	0.00045	6.00000
scaled y_3	2.4207	0.00045	5.51516

Table 4.22: Cost efficiency estimates using FF functional form without control for cross-country heterogeneities

	MO approach					BM approach				AGMM approach			
	FE	Pitt and Lee RE	Battese and Coelli RE	True FE	True RE	FE	Pitt and Lee RE	Battese and Coelli RE	True RE	FE	Pitt and Lee RE	Battese and Coelli RE	True RE
Whole Sample	0.3596	0.7184	0.7148	0.5238	0.9000	0.2488	0.4433	0.4465	0.9059	0.2698	0.4863	0.4584	0.9752
China	0.3769	0.7946	0.7885	0.5247	0.9014	0.2223	0.4164	0.4106	0.8992	0.2404	0.4620	0.4215	0.9752
Hong Kong	0.4235	0.7847	0.7747	0.5171	0.8979	0.2800	0.4936	0.4977	0.9078	0.3144	0.5557	0.5237	0.9754
India	0.3173	0.6804	0.6797	0.5324	0.9070	0.2168	0.4097	0.4133	0.9143	0.2408	0.4617	0.4340	0.9750
Indonesia	0.4164	0.6597	0.6537	0.5082	0.8993	0.3057	0.4657	0.4611	0.9014	0.3213	0.4887	0.4668	0.9752
Korea	0.2997	0.7288	0.7252	0.5273	0.8943	0.2192	0.4549	0.4582	0.9068	0.2433	0.5267	0.4915	0.9753
Malaysia	0.5424	0.9158	0.9152	0.5258	0.9117	0.4323	0.6511	0.6572	0.9121	0.4667	0.7042	0.6801	0.9756
Philippines	0.3320	0.6058	0.6043	0.5327	0.8939	0.2482	0.4167	0.4281	0.9000	0.2670	0.4492	0.4359	0.9752
Singapore	0.3515	0.9003	0.9092	0.5240	0.9124	0.2079	0.4502	0.4634	0.8909	0.2270	0.5091	0.4866	0.9752
Taiwan	0.3286	0.6961	0.6973	0.5245	0.8955	0.2288	0.4379	0.4483	0.9069	0.2520	0.4901	0.4686	0.9751
Thailand	0.1748	0.5706	0.5561	0.5104	0.8821	0.1161	0.2864	0.2963	0.9037	0.1032	0.2449	0.2106	0.9756

Table 4.23: Cost efficiency estimates using FF functional form with control for cross-country heterogeneities

	MO approach		BM approach		AGMM approach	
	Pitt and Lee RE	Battese and Coelli RE	Pitt and Lee RE	Battese and Coelli RE	Pitt and Lee RE	Battese and Coelli RE
Whole Sample	0.7490	0.7486	0.5919	0.6054	0.5939	0.6072
China	0.7935	0.8065	0.5667	0.6128	0.5737	0.6158
Hong Kong	0.7992	0.7760	0.6603	0.6286	0.6719	0.6422
India	0.8359	0.8327	0.7711	0.7466	0.7832	0.7683
Indonesia	0.6968	0.6932	0.5526	0.5635	0.5391	0.5543
Korea	0.6885	0.6848	0.5030	0.5316	0.5258	0.5563
Malaysia	0.8558	0.8685	0.6239	0.6749	0.6413	0.6888
Philippines	0.6302	0.6305	0.5518	0.5653	0.5466	0.5652
Singapore	0.9469	0.9436	0.7241	0.7234	0.7115	0.7180
Taiwan	0.6769	0.6791	0.4911	0.5158	0.5001	0.5211
Thailand	0.5851	0.5714	0.3554	0.3827	0.2775	0.2691

With the preferred Battese and Coelli (1992) model written in FF cost functional form and incorporating the cross-country heterogeneities, comparisons between these three approaches can be implemented (Table 4.24). Apparently, MO approach performs much better than the other two. 18 out of 24 Fourier terms are statistically significant. However, for BM approach, only 11 out of 40 Fourier terms are significant and for AGMM approach, only 3 out of 6 Fourier terms are significant[13]. Besides, MO approach has the largest likelihood function and coefficients of outputs and input prices are all statistically significant while in BM approach, coefficients of Y2 and Y3 is insignificant and in AGMM approach, coefficient of Y3 is insignificant.

In next stage, by focusing on MO approach, I attempt to test the FF functional form agaist the translog functional form. Based on the LR test, it seems that the FF functional form is the better candidate than the translog functional form as the LR ratio is 324.734 (the critical value for chi-square distribution with 24 degree of freedom is 42.98 at the

[13] Even with the same Fourier terms as used in MO approaches, BM approach still performs worse although 16 out of 24 Fourier terms are significant. The coefficient of Y3 is still insignificant. The same result applies for AGMM approach.

Table 4.24: Key estimates from BC model adopting three FF approaches with control for cross-country heterogeneities

Coefficients	MO approach	BM approach	AGMM approach
Outputs			
Y1	1.435	0.478	0.473
Y2	1.234	0.209†	0.186
Y3	0.673	0.159†	0.059†
Input prices			
W1	0.486	0.842	0.714
W2	0.729	0.685	0.157
Cross-country environmental variables			
Z1	0.073	0.133	0.133
Z2	-0.161	-0.275	-0.283
Z3	0.195	0.276	0.192
Z4	-0.045	-0.018†	-0.044
Z5	-0.073	-0.041†	-0.059
Z6	-0.035†	-0.143	-0.091
Z7	0.006†	-0.061	-0.073
Z8	-0.066	-0.032†	-0.038†
Technical progress			
Neutral	yes	yes	yes
Non-neutral	yes	no	no
6 Coefficients	4 significant	0 significant	1 significant
FF series terms	24 terms	40 terms	18 terms
First-order trigonometric terms	8 out of 10 are significant	2 out of 10 are significant	1 out of 6 are significant
Second-order trigonometric terms	10 out of 14 are significant	9 out of 30 are significant	2 out of 12 are significant
Maximum likelihood function			
LLF	1315.058	1178.877	1058.052

Notes: † denotes insignificance at the 10% level.

1% level and 33.20 at the 5% level). This finding coincides with Mitchell and Onvural (1996) in which the authors find strong evidence in favor of the FF functional form. I also test the FF functional form with only the first-order trigonometric Fourier terms against the one with the first- and second-order trigonometric Fourier terms. The LR ratio of 126.077 also suggests that the second-order trigonometric Fourier terms should be included in the cost function to generate more accurate results. However, when checking the monotonicity and concavity condition of the underlying cost function (see Table 4.25), the FF functional form passes the monotonicity test but fails the concavity

Table 4.25: Properties (Monotonicity, scale and concavity) of the FF cost function using MO approach

Monotonicity Property	**Elasticity**	**Parameters**	**Standard errors**	**Whole sample**
at the sample mean	ey1	1.435	0.199	99.8
at the sample mean	ey2	1.234	0.180	99.1
at the sample mean	ey3	0.673	0.56	84
at the sample mean	ew1	0.486	0.047	99.9
at the sample mean	ew2	0.729	0.073	78.2
Scale Property	**Scale Elasticity**	**Standard errors**	**t-value**	**Whole sample:**
at the sample mean	E = 3.342	0.3265	7.176 Reject H_0: E=1	100
Concavity Property	**Objective function**	**Principle Minors**	**Values**	**Whole sample**
at the sample mean	H(w)	First order	-0.274, -0.150	1.27
		Second order	-0.122	

condition both at the sample mean (as the second order principle minors are negative) and over the sample (only 1.27% sample points satisfy the concavity condition). Therefore, despite the sacrifice of flexibility for global approximations, the translog functional form provides parameter estimates from the regression of true cost function as defined in theory. Moreover, this result also suggests that the pursuit of globally fitted cost functions in banking efficiency studies may not rely on using more flexible functional forms that may not fit the data properly and may cause estimation problem when the sample data is relatively small. Instead, the resolution may lie in introducing additional explanatory variables to the cost function as a means of providing better description of production process.

4.6 Conclusion

To fill the literature gap of cross-country banking efficiency studies, this chapter examines and compares cost efficiency in ten Asian banking industries from 1998 to 2005. Unlike the previous cross-country studies, this work uses the panel data stochastic frontier approach. Compared with the cross-sectional framework, panel data approaches can relax the strict distributional assumption of the inefficiency term and most importantly generate the consistent inefficiency estimates.

In cross-country studies, differences in inefficiency may be attributed to not only the managerial ability of the banks but also partly to the different characteristics of countries. Excluding cross-country environmental variables may have underestimated efficiency scores. As seen from empirical results, when cross-country heterogeneities are considered, cost efficiency scores are higher than when they are not included.

From the preferred Battese and Coelli (1992) model with incorporation of cross-country heterogeneities, the overall cost efficiency in these Asian banking industries is 0.5897 with a decreasing trend, despite positive technical progress and slight economies of scale. It is possible that during the post-crisis financial reform and reconstruction of the banking system, attempts to reduce the large volume of NPLs and to improve asset quality led to a decline in the loan output with given input usage, resulting in a decrease in efficiency scores. Also, mergers and acquisition between healthy banks and troubled banks, forced by governments, may have led to a decline in the performance of commercial banks during the sample period. When comparing efficiency scores country by country, I find that India, Singapore, Malaysia and Hong Kong SAR seem to have the most efficient banking industries in Asia.

This chapter ends up with further discussions of two long-standing debates on the use of output and input specifications and the appropriate functional forms. Based on close examination of the existing studies and my own results, the intermediation approach seems to be the right choice when data only on the value of total deposits is available. The production approach and dual approach may be applicable if data of the number of total deposit accounts is available. However, to avoid the misleading policy implications and suggestions, certain statistical inferences and theoretical properties have to be checked before jumping to any conclusion. The argument on the use of functional form mainly concentrates on the flexibility of the underlying functions but neglects their credibility. It could always be the first priority to make sure that cost efficiency is truly estimated from the true cost function, even with the sacrifice of flexibility.

Chapter 5 Developing an Index Number Approach to Productivity Decomposition: with an Application to Asian Banking Industries

5.1 Introduction

In the past two decades, banking and financial systems have experienced dramatic changes and developments all over the world. On the one hand, banking deregulation, financial integration and merger and acquisition remark the extensive transformation of banking operational environment. On the other hand, driven by the technological innovation, banks are able to save costs in providing financial services and create a range of new products. Inspired by these developments, a substantial body of efficiency and productivity studies have been conducted on purpose to inform regulators and practitioners of banking sector performance in this rapidly changing environment and help governor review banking and financial regulation and supervision and assist bank managers to assess their managerial ability.

In the literature of performance evaluation studies, a large number of researchers focus on technical efficiency, cost efficiency and profit efficiency using either non-parametric or/and parametric frontier methodologies. Another strand of academic studies has focused on productivity measurement and its decomposition using either non-parametric or/and parametric frontier approaches. The former uses the Malmquist productivity index, first proposed by Caves *et al.* (1982), to measure the TFP change and then to decompose the productivity change into several sources. This Malmquist productivity index is calculated using DEA. The latter studies the magnitude of TFP change using econometric techniques, with earlier attempts focusing on the deterministic production frontier based on the assumption of single output and single input and more recent developments using dual approaches, such as cost function, distance function and profit function that allow for multiple outputs under a stochastic frontier framework that allows for random noise. Detailed discussions on various approaches of productivity measurement are provided in section 2.3.

The paradigm of parametric stochastic frontier approaches can be categorized into two main approaches, which are total differential approach and index number approach. Under total differential approach, productivity change and its determinants are measured by derivatives from total differentiating the production function, (or cost function, or profit function, or distance function) with respect to a time trend variable. Thus, productivity change is measured for the whole sample period, for example, measuring productivity growth from period t to t+5 as a whole. However, under index number approach productivity change is measured in discrete time. It allows one to measure and compare the trend of productivity change for the sub sample period, for example, measuring productivity growth for period t to t+1, t+1 to t+2, etc,. This feature is extremely desirable in empirical studies as it is always researchers' interest to evaluate the productivity change per annum throughout the whole sample and this effort clearly indicates the annual trend of productivity change and allows researchers to analyze the driven sources of productivity growth, stagnation or decline. So far, the only index number approach can be found is Orea (2002), in which the author develops a generalized Malmquist productivity index using a stochastic output oriented distance function. However, because my interest here is to evaluate the TFP change of Asian banking industries from a cost perspective, it is ideal to adopt an index number approach using cost function. This desire is fulfilled by developing an index number counterpart of Bauer's (1990) total differential approach that measures the TFP change using a stochastic translog cost function. The theoretical framework of this cost-based index number approach has only been proposed in Coelli *et al.* (2003:41). However, their derivation of cost efficiency change and scale effect change components is inappropriate, as well as the sign of allocative efficiency change component.

The literature of productivity studies in banking sectors has been dominated by studies focusing on the banking sector in the US and European Union, most of which has employed a non-parametric Malmquist productivity index. Limited attempts have tried to measure the productivity change in Asia. Asian banking industries, for instance, Hong Kong, Singapore, China, India and Korea etc., have changed and developed a lot since the 1990s. Especially after the Asian financial crisis, Asian banking and financial sectors have experienced a substantial reform with attempts to deepen the banking deregulation, strengthen banking supervision and credit monitoring. Financial integration in Asian countries is progressing and a common Asian banking and financial market marks the trend. However, so far no attempts have been carried out to measure and compare the productivity change in main Asian countries and regions on purpose to inform regulator and bankers on banking performance in this dramatic changing banking environment. Therefore, this chapter intends to fill this literature gap by measuring the TFP change and its sources in main Asian banking industries by adopting parametric stochastic cost frontier approach.

The rest of this chapter is organized as follows. Section 5.2 provides detailed review of the existing productivity literature in the banking sector that intrigues my interests. Section 5.3 introduces the detailed derivation of Bauer's (1990) total differential approach in the case of multiple outputs. In Bauer's (1990) original paper, although he provides both single-output and multiproduct version of TFP change and its decomposition, he only provides the mathematical derivation in the case of single output cost frontier. Duplicating the mathematical derivation of decomposition of TFP change in the case of multiple outputs will help build up the foundation for developing the index number counterpart of Bauer's approach in section 5.4. Section 5.5 provides the empirical results of TFP change in these Asian banking industries. Section 5.6 concludes.

5.2. Literature Review on banking productivity studies

A survey has been conducted and presented in Table 5.1. Literature of measurement of productivity change in banking has been dominated by the utilization of Malmquist productivity index, most of which are calculated through non-parametric approach DEA. One of the first studies to measure productivity change in the banking industry was by Berg *et al.* (1992). By employing Malmquist productivity indices, they find evidence of positive deregulation effects on productivity as the initial productivity shrinking is followed by rapid growth in Norwegian banking sector during 1980-89. Grifell-Tatjé and Lovell (1996) investigates the TFP change in Spanish saving banks over the 1986-91. Their results from Malmquist productivity index indicate a productivity decline due to the reason of neither branching nor mergers over the sample period. Grifell-Tatjé and Lovell (1997) use the generalized Malmquist productivity index to study the sources of productivity change in Spanish banking. They find that commercial banks had a slightly slower rate of productivity growth, but a slightly higher rate of potential productivity growth. Noulas (1997) studies the productivity growth of the Hellenic banking industry by using Malmquist productivity index in 1991 and 1992. The results indicate that although productivity has increased for both state and private banks, the sources of this growth are different with state-owned banks' productivity growth coming from technological progress but private banks' from increased efficiency. Wheelock and Wilson (1999) constructed Malmquist productivity index from distance function to measure efficiency and productivity change in the US banks from 1984-93. Their results suggest that productivity increases on average for those banks with asset size over $300million and productivity declines on average for those with asset size less than $300 million. Avkiran (2000) look into Australian banking industry using Malmquist type index in a deregulated period 1986-95. Their principal findings indicate an overall increase in productivity mainly driven more by technical progress than technical efficiency. Devaney and Weber (2000) estimate the Malmquist productivity index for the US rural banking sector over the period 1990-93. Their results suggest that rural bank's productivity growth for the three-year period is 11%, attributed to technological change rather than pure technical change or scale change. Alam (2001) uses a similar approach to investigate productivity change in large US commercial banks over 1980s and find that despite a fall in productivity in 1985, all banks make tremendous gains in productivity in 1980s. Mukherjee *et al.* (2001)

Table 5.1: Survey of productivity studies in banking sector

Authors	Applied countries	Period	Methodology	Decomposition	International comparison
Alam (2001)	US	1980-89	Malmquist (DEA)	TC, EC	No
Avkiran (2000)	Australia	1986-95	Malmquist (DEA)	TC, EC	No
Berg *et al.* (1992)	Norway	1980-89	Malmquist (DEA)		No
Berger and Mester (1999)	US	1991-97	Parametric BM approaches	BM decomposition	No
Berger and Mester (2003)	US	1991-97	Parametric BM approaches	BM decomposition	No
Canhoto and Dermine (2003)	Portugal	1990-95	Malmquist (DEA)	TC, EC	No
Casu *et al.* (2004)	5 EU countries	1994-00	Malmquist (DEA), Parametric BM approaches	TC, EC, SEC; BM decomposition	Yes
Devaney and Weber (2000)	US	1990-93	Malmquist (DEA)	TC, EC, SEC	No
Fukuyama (1995)	Japan	1989-91	Malmquist (DEA)	TC, EC	No
Fukuyama and Weber (2002)	Japan	1992-96	Malmquist (DEA)	TC, EC	No
Gilbert and Wilson (1998)	Korea	1980-84	Malmquist (DEA) and bootstrap	TC, EC, SEC	No
Grifell-Tatjé and Lovell (1996)	Spain	1986-91	Malmquist (DEA)	TC, EC	No
Grifell-Tatjé and Lovell (1997)	Spain	1986-93	Malmquist (DEA)	TC, EC, SEC	No
Humphrey (1993)	US	1977-88	Parametric	TC	No
Isik and Hassan (2003a)	Turkey	1981-90	Malmquist (DEA)	TC, EC, SEC	No
Isik and Hassan (2003b)	Turkey	1992-96	Malmquist (DEA)	TC, EC, SEC	No
Kumbhakar *et al.* (2001)	Spain	1986-95	Parametric TFP index	TC, EC	No
Kumbhakar and Lozano-Vivas (2005)	Spain	1987-00	Parametric TFP index (SFA)	TC, SEC, AEOC, TCz	No
Kumbhakar and Wang (2007)	China	1993-02	Parametric TFP index (SFA)	TC, EC, SEC, TCz	No

Table 5.1: Survey of productivity studies in banking sector (continued)

Authors	Applied countries	Period	Methodology	Decomposition	International comparison
Mukherjee *et al.* (2001)	US	1984-90	Malmquist (DEA)	TC, EC, SEC	No
Murillo-Melchor *et al.* (2005)	14 EU countries	1995-01	Malmquist (DEA) and bootstrap	TC, EC	Yes
Noulas (1997)	Greece	1991-92	Malmquist (DEA)	TC, EC	No
Orea (2002)	Spain	1985-98	Malmquist (SFA)	TC, EC, SEC	No
Stiroh (2000)	US	1991-97	Parametric BM approaches	BM decomposition	No
Tortosa-Ausina *et al.* (2008)	Spain	1992-98	Malmquist (DEA) and bootstrap	TC, EC	No
Tsionas *et al.* (2003)	Greece	1993-98	Malmquist (DEA)	TC, EC, SEC	No
Wheelock and Wilson (1999)	US	1984-93	Malmquist (DEA)	TC, EC	No

explore productivity growth for large US commercial banks over the initial post-deregulation period from 1984 to 1990. By using DEA-type Malmquist productivity index, they find the overall productivity growth over the period but except for 1984-85 and 1988-89. Canhoto and Dermine (2003) study the impact of deregulation on Portuguese commercial banks' performance using Malmquist index and they find evidence of positive productivity growth, mainly driven by the technical progress. Isik and Hassan (2003a) utilize a DEA-type Malmquist productivity index to examine productivity growth in Turkish commercial banks during the deregulation of financial markets between 1981 and 1990. They find that all forms of Turkish banks, although in different magnitudes, have recorded significant productivity gains driven mostly by efficiency increases rather than technical progress. Isik and Hassan (2003b) use the same approach as Isik and Hassan (2003a) to investigate the impact of 1994 Turkish financial crisis on productivity of Turkish banking sector. Consistent with the descriptive analysis, their empirical results suggest a 17% decline to productivity that is mainly attributed to technical regress (10%). Tsionas *et al.* (2003) estimate economic efficiency, TFP change, and technical change of the Greek banking system over the period 1993-98. From their Malmquist productivity index, positive but not substantial

TFP change is observed in Greek banking system.

Three studies (Gilbert and Wilson, 1998; Tortosa-Ausina *et al.*, 2008 and Murillo-Melchor *et al.*, 2005) have used bootstrapping techniques so as to construct the confidence intervals for the efficiency scores and productivity indices to address the main drawback of DEA-type Malmquist productivity index. The first study was conducted by Gilbert and Wilson (1998) that analyzed the effects of deregulation on the productivity of Korean banks over 1980-84. They found that Korean banks had experienced substantial changes in productivity during the privatization and deregulation process. Tortosa-Ausina *et al.* (2008) explore productivity growth and productive efficiency for Spanish savings banks over the initial post-deregulation period 1992–98 using DEA and bootstrapping techniques. Results suggest that productivity growth has occurred, mainly due to improvement in production possibilities. Murillo-Melchor *et al.* (2005) analyze productivity growth for European banks over 1995-2001 with results indicating the positive productivity growth in European banks.

A number of studies have used various econometric model specifications to estimate the TFP change. By estimating a cost function, Humphrey (1993) find negative productivity growth in the US banks during 1977-89. Kumbhakar *et al.* (2001) examine the impact of regulatory reform on the performance of Spanish savings bank for the period of 1986-95 by using a parametric decomposition of TFP change from profit function. Empirical results suggest the declining technical efficiency but also find evidence of technical progress and positive productivity growth. Orea (2002) derives a parametric generalized Malmquist productivity index from the output distance function. In their application to Spanish savings banks over 1985-98, positive productivity growth is found, mainly attributed to technical progress and positive effects of returns to scale. Kumbhakar and Lozano-Vivas (2005) develop a parametric decomposition to TFP change by using the cost function in a continuous time and apply it to the Spanish commercial and savings banks from 1987 to 2000. They find that deregulation contributed positively to the productivity growth for both savings and commercial banks.

There are other parametric studies measuring the productivity change without introducing the time trend in the respective econometric model. Berger and Mester

(1999, 2003) introduce a decomposition of total cost changes into a portion due to changes in business conditions and a portion due to changes in productivity. They find cost productivity deterioration despite substantial profit productivity improvement from 1991–97. Stiroh (2000) employs various econometric techniques, including the cost decomposition suggested by Berger and Mester (1999, 2003) and finds productivity growth of about 0.4% and these results are found to be consistent across different methodologies.

Compared to productivity analysis of individual banking sector, literature of international comparison of productivity studies is limited. Casu *et al.* (2004) compare parametric and non-parametric estimates of productivity change in European banking between 1994 and 2000. The two approaches generally yield similar findings suggesting that productivity growth has mainly been brought by improvements in technological change.

Although the current productivity literature has been dominated by studies in the US and European banking sector, a limited number of scholars have applied various methodologies to measure the TFP change in Asian banking industries and its potential sources from decomposition. Fukuyama (1995) uses non-parametric Malmquist productivity index to measure the productivity growth in Japanese banking industry from 1989-1991. Using panel data of Japanese banks operating from period 1992-96, Fukuyama and Weber (2002) measure the productivity growth and its decomposition from a DEA-type Malmquist productivity index and find that Japanese banks have experienced productivity decline averaging 2% per year and could have used only 78-93% of actual inputs if they had chosen the revenue maximizing output mix. As mentioned earlier, Gilbert and Wilson (1998), who adopt the bootstrap technique and Malmquist productivity index, measure the productivity change in Korean banking industry. Kumbhakar and Wang (2007) analyze the impact of banking reforms on efficiency and TFP change in Chinese banking industry for 1993-2002. The overall TFP growth is 4.4% per annum with joint-equity banks experiencing much higher growth in TFP compared to wholly state-owned banks.

However, there is no study that measures and compares the productivity change between main Asian economies. Moreover, from methodological perspective, no study constructs the index number parametric decomposition of TFP change derived from cost function. Furthermore, most TFP change studies are conducted through DEA-type Malmquist productivity index, which normally decompose the TFP change into technical change, technical efficiency change and effect of returns to scale. However, few attempts in banking studies have tried to derive further decompositions to obtain allocative efficiency changes, which are mainly attributed to input prices effects from the bias of using the real cost share weights instead of the optimal cost share weights and output price effects from the bias of using the nonmarginal cost pricing, which is using the real revenue share weights instead of the optimal revenue share weights. To fill this literature gap, a parametric five-way decomposition of TFP change is derived using panel data stochastic cost frontier approach with an application to Asian banking industries.

5.3. Derivation of Bauer's (1990) approach in the case of multiproduct firms

A firm i is said to produce J outputs y_{jt} $\left(\mathrm{j}=1,...,\mathrm{J}\right)$ using M inputs x_{mt} $\left(\mathrm{m}=1,...,\mathrm{M}\right)$ at time t (t=1,…,T). Suppose the output and input price is p_{jt} nd w_{mt} for the jth output and the mth input at time t, respectively. Follow Bauer's (1990), TFP growth is measured as the net change in outputs to inputs, written in the case of multiple outputs Divisia index of TFP change as

$$T\dot{F}P = \sum_{j} r_j \cdot \dot{y}_j - \sum_{m} s_m \cdot \dot{x}_m \qquad [5.1]$$

where

$T\dot{F}P$, $\dot{y}_j$ and $\dot{x}_m$ are the derivatives of the logarithm term of TFP change, the jth output and the mth input with respect to time.

r_j is the actual revenue share of output y_j to the total revenue as $r_j = p_j y_j \Big/ \sum_j p_j y_j$.

s_m is the actual cost share of input x_m to the total cost, measured as $s_m = w_m x_m \Big/ \sum_m w_m x_m$.

As to maximize production of outputs, the firm intends to minimize the cost. Therefore, the cost function can be defined as

$$C = C(\mathbf{y}, \mathbf{w}, t) = \min\left(\sum_m x_m w_m : (x, y) \in T \right) \qquad [5.2]$$

and the actual cost observed for the firm is

$$C_o = \sum_m x_m w_m \qquad [5.3]$$

Since cost efficiency reflects how the firm performs against the best practice represented by the cost frontier, it can be measured as the deviation of the actual observed cost from the estimated efficient cost frontier.

$$CE = \frac{C(\mathbf{y}, \mathbf{w}, t)}{C_o} \qquad \text{where } CE \le 1 \qquad [5.4]$$

Taking logarithm on both sides of equation [5.4] and total differentiating with respect to time t, we can have

$$\frac{\partial \ln CE}{\partial t} = \sum_j \frac{\partial \ln C(\mathbf{y},\mathbf{w},t)}{\partial \ln y_j} \cdot \frac{\partial \ln y_j}{\partial t} + \sum_m \frac{\partial \ln C(\mathbf{y},\mathbf{w},t)}{\partial \ln w_m} \cdot \frac{\partial \ln w_m}{\partial t} + \frac{\partial \ln C(\mathbf{y},\mathbf{w},t)}{\partial t} - \frac{\partial \ln C_o}{\partial t}$$

[5.5]

Rewrite equation [5.5] as

$$\dot{CE} = \sum_j ey_j \cdot \dot{y}_j + \sum_m s_m^* \cdot \dot{w}_m + \dot{C}(\mathbf{y},\mathbf{w},t) - \dot{C}_o \qquad [5.6]$$

where

$\dot{CE}$, $\dot{C}_0$, $\dot{w}_m$ and $\dot{x}_m$ is also defined as derivatives of the logarithm terms of respective variables with respect to time.

ey_j measures the elasticity for the jth output as $ey_j = \partial \ln C(\mathbf{y},\mathbf{w},t) / \partial \ln y_j$.

s_m^* is input prices elasticity, $s_m^* = ew_m = \partial \ln C(\mathbf{y},\mathbf{w},t) / \partial \ln w_m$, that measures the optimal cost share, also known as share equations in regard to Shephard lemma.

Note the following

$$\frac{\partial \ln C_o}{\partial t} = \frac{\partial C_o}{\partial t} \cdot \frac{1}{C_o} = \frac{1}{C_o} \left[\sum_m x_m \cdot \frac{\partial w_m}{\partial t} + \sum_m w_m \cdot \frac{\partial x_m}{\partial t} \right] \qquad [5.7]$$

which can be rewritten as

$$
\begin{aligned}
\dot{C}_0 &= \sum_m \frac{x_m w_m}{C_o} \cdot \frac{1}{w_m} \cdot \frac{\partial w_m}{\partial t} + \sum_m \frac{w_m x_m}{C_o} \cdot \frac{1}{x_m} \cdot \frac{\partial x_m}{\partial t} \\
&= \sum_m \frac{x_m w_m}{C_o} \cdot \dot{w}_m + \sum_m \frac{w_m x_m}{C_o} \cdot \dot{x}_m \\
&= \sum_m s_m \cdot \dot{w}_m + \sum_m s_m \cdot \dot{x}_m
\end{aligned}
\quad [5.8]
$$

where s_m is the actual cost share defined earlier.

Recall the definition of TFP change for the multiple outputs firm in [5.1]. By inserting [5.6] and [5.8] into [5.1], one can obtain the multiproduct version of parametric decomposition of the TFP change, written as

$$
T\dot{F}P = (1-E)\cdot \sum_j r_j^* \cdot \dot{y}_j + \dot{C}E - \dot{C}(\mathbf{y}, \mathbf{w}, t) + \sum_m \left(s_m - s_m^*\right)\cdot \dot{w}_m + \sum_j \left(r_j - r_j^*\right)\cdot \dot{y}_j \quad [5.9]
$$

where $E = ey_j = \partial \ln C(\mathbf{y}, \mathbf{w}, t) / \partial \ln y_j$ is the sum of output elasticities;

$r_j^* = ey_j / E$ is measured as the optimal revenue share.

This expression of TFP decomposition [5.9] is exactly the same as Bauer's (1990) multiproduct version but with different notation. It identifies five potential sources that drive productivity growth.

1) $(1-E)\sum_j r_j^* \cdot \dot{y}_j$ is the term of scale effect change. In the cost perspective, the firm is said to benefit from economies of scale if the sum of output elasticities is less than one. This will be reflected as a positive sign in $(1-E)\sum_j r_j^* \cdot \dot{y}_j$;
2) $\dot{C}E$ measures cost efficiency change, which can be further decomposed into technical efficiency and allocative efficiency change;
3) $-\dot{C}(\mathbf{y}, \mathbf{w}, t)$ is the technological change. In the presence of technical progress, the firm is benefiting from using advanced technology in the production process. Its cost should be decreasing over time, which will be reflected as a negative sign in $\dot{C}(\mathbf{y}, \mathbf{w}, t)$ Therefore, $-\dot{C}(\mathbf{y}, \mathbf{w}, t)$ indeed indicates the hypothesis of technical progress;

4) $\sum_m (s_m - s_m^*) \cdot \dot{w}_m$ is the input price effect from the wrong mix of input quantities, measured as the allocative inefficiency by the firm. If the firm is allocatively efficient, then $s_m = s_m^*$ and the price effect term will be equal to zero.

5) $\sum_j (r_j - r_j^*) \cdot \dot{y}_j$ is the effect of bias in not using marginal cost pricing on the observed measure of TFP change. This measurement requires information on the output prices.

However, as explained in the introduction of this chapter, this five-way decomposition from [5.9] is defined on continuous time and it can hardly be applied to empirical productivity studies as most of which are implemented on discrete time to enable comparisons in the year basis. To make it applicable, this total differential approach has to be converted to a discrete time formulation.

5.4 Derivation of index number counterpart of Bauer's approach

Suppose the firm's cost function defined in [5.2] is represented by the translog cost function. Note that this translog cost function can be regarded as a quadratic function in the variables y^t, w^t and t. Hence, it is possible to apply Diewert's (1976) Quadratic Identity Lemma. Based on that, changes in cost function from period t to t+1 can be written as:

$$\ln\left[\frac{C(t+1)}{C(t)}\right] = \frac{1}{2}\sum_j \left[ey_j^{t+1} + ey_j^t\right] \cdot \ln\left(\frac{y_j^{t+1}}{y_j^t}\right) + \frac{1}{2}\sum_m \left[s_m^{*,t+1} + s_m^{*,t}\right] \cdot \ln\left(\frac{w_m^{t+1}}{w_m^t}\right) + \frac{1}{2}\left[\frac{\partial \ln C(t+1)}{\partial t} + \frac{\partial \ln C(t)}{\partial t}\right]$$

[5.10]

Where $C(t)$ is short for $C(\mathbf{y}^t, \mathbf{w}^t, t)$;

ey_j^t is the elasticity for the jth ouput at time t;

$s_m^{*,t}$ is the optimal cost share for the mth input at time t;

From Törnqvist approximation[14], [5.8] can also be written in the context of discrete time as

$$\ln\left(\frac{C_o^{t+1}}{C_o^t}\right)=\frac{1}{2}\sum_m\left[s_m^{t+1}+s_m^t\right]\cdot\ln\left(\frac{x_m^{t+1}}{x_m^t}\right)+\frac{1}{2}\sum_m\left[s_m^{t+1}+s_m^t\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)$$

[5.11]

where s_m^t is the actual observed cost share for the mth input at time t.

Note that the logarithm term of cost efficiency level at time period t and t+1 as shown in [5.4] can be written as

$$\ln CE^t=\ln C(t)-\ln C_o^t \qquad [5.12]$$

$$\ln CE^{t+1}=\ln C(t+1)-\ln C_o^{t+1} \qquad [5.13]$$

Then from [5.12] and [5.13], it is quite straightforward to have

$$\ln\left(\frac{CE^{t+1}}{CE^t}\right)=\ln\left[\frac{C(t+1)}{C(t)}\right]-\ln\left(\frac{C_o^{t+1}}{C_o^t}\right) \qquad [5.14]$$

Putting [5.10] and [5.11] into [5.14], one would have

$$\ln\left(\frac{CE^{t+1}}{CE^t}\right)=\frac{1}{2}\sum_j\left[ey_j^{t+1}+ey_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)+\frac{1}{2}\sum_m\left[s_m^{*,t+1}+s_m^{*,t}\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)$$
$$+\frac{1}{2}\left[\frac{\partial\ln C(t+1)}{\partial t}+\frac{\partial\ln C(t)}{\partial t}\right]-\frac{1}{2}\sum_m\left[s_m^{t+1}+s_m^t\right]\cdot\ln\left(\frac{x_m^{t+1}}{x_m^t}\right)-\frac{1}{2}\sum_m\left[s_m^{t+1}+s_m^t\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)$$

[5.15]

Now let us define the TFP change in the context of discrete time. Following Törnqvist approximation, definition of TFP change shown in [5.1] can be rewritten as

$$\ln\left[\frac{TFP(t+1)}{TFP(t)}\right]=\frac{1}{2}\sum_j\left[r_j^{t+1}+r_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)-\frac{1}{2}\sum_m\left[s_m^{t+1}+s_m^t\right]\cdot\ln\left(\frac{x_m^{t+1}}{x_m^t}\right)$$

[5.16]

---

14 Törnqvist approximation shows $\theta_j\dot{y}_j=\Delta\theta_j\ln y_j=\frac{1}{2}\left[\theta_{jt}+\theta_{j,t-1}\right]\cdot\ln\left(y_{jt}/y_{j,t-1}\right)$

where $TFP(t)$ is short for $TFP(\mathbf{y}^t, \mathbf{w}^t, t)$.

r_j^t is the observed revenue share for the jth output at time t.

s_m^t is the observed cost share for the mth input at time t.

Here TFP change is defined as a weighted index of output change minus a weighted index of input change. The weights are the actual revenue and cost shares as we define before.

Under index number approaches, the TFP index should satisfy certain properties (see Fried *et al.* 2008: 62). There is a general consensus that four desirable properties should be satisfied: identity, monotonicity, separability, and proportionality. The identity property implies that if inputs and outputs do not change from period t to t+1, the TFP change equal to one. The monotonicity property requires that a TFP index is constructed with higher output and lower input usage indicating the improvement in productivity. Separability implies that a TFP index is interpreted in the same way as in the single-output single-input case, for example, in the multiple-outputs multiple-inputs case, the aggregated output growth rate only relies on the output data and the aggregated input growth rate only depends on the input data. Therefore, if the technology is separable in outputs and inputs, the TFP index has the desirable property. The proportionality property suggests that the weights in the output and input growth indices should add to one. Apparently, the TFP index defined in [5.16] satisfies above four properties. The structure of the TFP index ensures the satisfaction of identity and monotonicity property. Since the output (input) growth rate only depends on the output (input) level and output (input) price information, the TFP index satisfies the separability condition. Moreover, because the weights in output and input growth indices are the actual observed revenue and cost shares, both of which sums to one, the proportionality properties is fulfilled.

Finally, the index number approach of decomposition of the TFP change can be obtained by substituting [5.15] into [5.16] (see **Appendix 3** for the detailed transformation on this final step). This is

$$\ln\left[\frac{TFP(t+1)}{TFP(t)}\right]$$
$$=\frac{1}{2}\sum_j\left[r_j^{*,t+1}\left(1-E^{t+1}\right)+r_j^{*,t}\left(1-E^t\right)\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)+\ln\left(\frac{CE^{t+1}}{CE^t}\right)-\frac{1}{2}\left[\frac{\partial\ln C(t+1)}{\partial t}+\frac{\partial\ln C(t)}{\partial t}\right]$$
$$+\frac{1}{2}\sum_m\left[\left(s_m^{t+1}-s_m^{*,t+1}\right)+\left(s_m^t-s_m^{*,t}\right)\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)+\frac{1}{2}\sum_j\left[\left(r_j^{t+1}-r_j^{*,t+1}\right)+\left(r_j^t-r_j^{*,t}\right)\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)$$
[5.17]

The expression [5.17] provides the discrete parametric decomposition of the TFP change using cost function:

1) $\frac{1}{2}\sum_j\left[r_j^{*,t+1}\left(1-E^{t+1}\right)+r_j^{*,t}\left(1-E^t\right)\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)$ is the scale effect change. In the presence of economies of scale (increasing returns to scale), the sum of output elasticities should be less than one, as $E^t<1$. This provides a positive sign of scale effect change term, indicating productivity growth.
2) $\ln\left(CE^{t+1}/CE^t\right)$ is the cost efficiency change;
3) $-\frac{1}{2}\left[\frac{\partial\ln C(t+1)}{\partial t}+\frac{\partial\ln C(t)}{\partial t}\right]$ is the technical change. Technical progress corresponds to the negative sign of elasticity of the cost function with respect to time;
4) $\frac{1}{2}\sum_m\left[\left(s_m^{t+1}-s_m^{*,t+1}\right)+\left(s_m^t-s_m^{*,t}\right)\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)$ is the input price effects from bias of using the real cost share weights instead of the optimal cost share weights. If the firm is allocatively efficient, $s_m^t=s_m^{*,t}$, then this term will equal to zero;
5) $\frac{1}{2}\sum_j\left[\left(r_j^{t+1}-r_j^{*,t+1}\right)+\left(r_j^t-r_j^{*,t}\right)\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)$ is the output price effects from bias of using the nonmarginal cost pricing, which is using the real revenue share weights instead of the optimal revenue share weights.

The only other version of the cost-based index number decomposition of the TFP change is provided in Coelli *et al.* (2003:41) and applied in Rungsuriyawiboon and Coelli (2007). Based on multiple output cost frontier, without provision of detailed derivation, Coelli *et al.* (2003) also develop a five-way decomposition of the TFP change from period 0 and period 1, written in their notations as

$$\ln\left(TFP_{n1}/TFP_{n0}\right)=\ln\left(CE_{n0}/CE_{n1}\right)-0.5\left[\left(\partial C_{no}/\partial t\right)+\left(\partial C_{n1}/\partial t\right)\right]$$
$$+0.5\sum_{i=1}^{M}\left[\left(SF_{n0}\varepsilon_{in0}+SF_{n0}\varepsilon_{in0}\right)\left(y_{in1}-y_{in0}\right)\right]$$

[5.18]

where y_{in0}, in their notation, is the logarithm of output, SF_{nt} is the scale factor that is termed as $SF_{nt}=\left(\varepsilon_{nt}-1\right)/\varepsilon_{nt}$, where $\varepsilon_{nt}=\sum_{i=1}^{K}\varepsilon_{\text{int}}$. ε_{int} is the output elasticity for the ith output.

If output prices information is available, a further source of allocative efficiency change can be contributed to the TFP change. The allocative efficiency change can be further decomposed into input mix allocative efficiency change and output mix allocative efficiency change. That is, written in their notations,

$$AEC=0.5\sum_{i=1}^{K}\left\{\left[\left(\kappa_{in1}-s_{in1}\right)+\left(\kappa_{in0}-s_{in0}\right)\right]\cdot\left(w_{in1}-w_{in0}\right)\right\}$$
$$+0.5\sum_{j=1}^{M}\left\{\left[\left(\pi_{jn1}-r_{jn1}\right)+\left(\pi_{jn0}-r_{jn0}\right)\right]\cdot\left(y_{jn1}-y_{jn0}\right)\right\}$$

[5.19]

where in their notation, w_{in0} is the logarithm term of input price, κ_{in1} and s_{in1} are the optimal and observed cost share at period 1, respectively. π_{jn1} and r_{jn1} are the optimal and observed revenue share at period 1, respectively.

Combining [5.18] and [5.19] provides the Coelli *et al.* (2003) version (CEPT version) of TFP change, shown in equation [5.20].

$$\ln\left(TFP_{n1}/TFP_{n0}\right)$$
$$=\ln\left(CE_{n0}/CE_{n1}\right)-0.5\left[\left(\partial C_{no}/\partial t\right)+\left(\partial C_{n1}/\partial t\right)\right]+0.5\sum_{i=1}^{M}\left[\left(SF_{n0}\varepsilon_{in0}+SF_{n0}\varepsilon_{in0}\right)\left(y_{n1}-y_{n0}\right)\right]$$
$$+0.5\sum_{i=1}^{K}\left\{\left[\left(\kappa_{in1}-s_{in1}\right)+\left(\kappa_{in0}-s_{in0}\right)\right]\cdot\left(w_{in1}-w_{in0}\right)\right\}$$
$$+0.5\sum_{j=1}^{M}\left\{\left[\left(\pi_{jn1}-r_{jn1}\right)+\left(\pi_{jn0}-r_{jn0}\right)\right]\cdot\left(y_{jn1}-y_{jn0}\right)\right\}$$

[5.20]

in which,

1). $\ln\left(CE_{n0}/CE_{n1}\right)$ captures the cost efficiency change term;

2). $-0.5[(\partial C_{no}/\partial t)+(\partial C_{n1}/\partial t)]$is the technical change term;

3). $0.5\sum_{i=1}^{M}[(SF_{n0}\varepsilon_{in0}+SF_{n0}\varepsilon_{in0})(y_{n1}-y_{n0})]$measures the scale effect change,

4). $0.5\sum_{i=1}^{K}\{[(\kappa_{in1}-s_{in1})+(\kappa_{in0}-s_{in0})]\cdot(w_{in1}-w_{in0})\}$is the input mix allocative efficiency change;

5). $0.5\sum_{j=1}^{M}\{[(\pi_{jn1}-r_{jn1})+(\pi_{jn0}-r_{jn0})]\cdot(y_{jn1}-y_{jn0})\}$is the output mix allocative efficiency change.

To compare the CEPT version with our version of TFP decomposition, it is clear that only technical change term is identical. Their cost efficiency change is measured as the logarithm ratio of cost efficiency at period 0 to the one at period 1, which is incorrectly calculated and should actually be the logarithm ratio of cost efficiency at period 1 to period 0 when TFP change is measured from period 0 to 1. Hence, it can be concluded that my measure of cost efficiency change is the correct one. Regarding the term of scale effect change, apparently, the CEPT version is constructed incorrectly. First of all, the weight term $(SF_{n0}\varepsilon_{in0}+SF_{n0}\varepsilon_{in0})$is either wrongly printed or miscalculated and the right formulation should be$(SF_{n1}\varepsilon_{in1}+SF_{n0}\varepsilon_{in0})$. Secondly, even if the weight term is constructed as$(SF_{n1}\varepsilon_{in1}+SF_{n0}\varepsilon_{in0})$, it is still an inappropriate measure of scale effect change. Compared to my version, in which scale effect change is positive if the firm is benefiting from economies of scale. However, in the CEPT version, it is a negative value, which usually suggests diseconomies of scale. Therefore, it can also be concluded that my measure of scale effect change is correct. Moreover, the CEPT version of allocative efficiency change term is also misleading since it measures the improvement of input and output mix allocative efficiency as a negative value, which usually indicates a deterioration of allocative efficiency. Thus, again, my version of allocative efficiency change is correct. To summarize, the CEPT version of decomposition of TFP change using the multiple outputs cost frontier is incorrectly constructed and employing it to measure the TFP change and identify its potential sources will generate incorrect results that actually provide the opposite image of the true TFP change.

Compared with the traditional approach, known as Malmquist productivity index, this cost-based TFP index provides more fruitful and economic meaningful decomposition. The first three parts (technical change, cost efficiency change and scale effect change) provide the similar decomposition to TFP change as seen in most productivity literature. However, different from this traditional Malmquist productivity index, my approach offers two more parts - input price effects and output price effects to reflect the allocative efficiency change. But unfortunately, in banking efficiency and productivity studies, it is difficult to get information on output prices, sometimes even input prices information. This partly explains why the Malmquist productivity index approach dominates the literature as it doesn't need information on input and output prices. However, since information on input prices is available, an input mix allocative efficiency change can be calculated in TFP change and empirical results indicate that allocative efficiency change is a very important source to explain the TFP change.

To summarize, built on my cost-based TFP index, the TFP change can be decomposed to five components, which are technical change, cost efficiency change, scale effect change, input mix allocative efficiency change and output mix allocative efficiency change. In the empirical application, output mix allocative efficiency change is left out because of the absence of data in output prices, then output mix allocative efficiency change is equal to zero by construction. This corresponds to an assumption of complete allocative efficiency in output prices since $r_j^t = r_j^{*,t}$. This corresponds to assuming that imperfect competition is absent from the markets for bank outputs. Consequently, there is a potential further source of TFP which could arise if output market becomes more competitive over the sample period. It cannot be measured at present because output prices data are unavailable. Therefore, the TFP change is measured as sum of the first four components these four components are calculated using the estimated coefficients from panel data stochastic cost frontier. Follow the model specification in last chapter, in which the environmental variables are assumed to influence the production technology, Battese and Coelli (1992) with incorporating cross-country heterogeneities is used here due to its flexibility of allowing efficiency to vary over time. The detailed model specification is described in chapter 4.

5.5 Empirical application

5.5.1 Model specification and empirical estimates

The same model specification of translog cost frontier shown in [4.7] to [4.9] is adopted here. Same parameter estimates and cost efficiency scores obtained from estimating Battese and Coelli (1992) model with incorporation of cross-country heterogeneities, as reported in Table 4.5, are used to calculate the TFP change and its decomposition.

5.5.2 Overall TFP change and its decomposition

TFP change is calculated from the parameter estimates reported in Table 4.5 and then decomposed to efficiency change, technical change, scale effect change and input mix allocative efficiency change. Results of average TFP change and its decomposition is summarized in Table 5.2 and plotted in Figure 5.1. Two sets of TFP change are calculated. The first set of TFP change, labeled as TFPC1 in Table 5.2, is calculated as the sum of efficiency change, technical change and scale effect change. This measurement of TFP change is similar to the Malmquist productivity index adopted in most productivity studies. The second set of TFP change, labeled as TFPC in Table 5.2, considers the contribution of allocative efficiency change in productivity growth, which was avoided or neglected in previous productivity literature. It is calculated as the sum of TFPC1 and input mix allocative efficiency change.

Table 5.2: Average TFP change and its components

Years	EC	TC	SEC	TFPC1	AEIC	TFPC
1998-99	-0.024	0.029	0.009	0.014	-0.021	-0.007
1999-00	-0.024	0.026	0.011	0.013	0.005	0.019
2000-01	-0.025	0.023	0.010	0.009	-0.004	0.005
2001-02	-0.026	0.021	0.012	0.007	0.018	0.024
2002-03	-0.027	0.018	0.013	0.004	0.013	0.018
2003-04	-0.028	0.015	0.013	0.000	0.025	0.026
2004-05	-0.029	0.014	0.013	-0.002	0.001	-0.001
1998-05	-0.026	0.021	0.012	0.006	0.005	0.012

Notes: EC=efficiency change, TC=technical change, SEC=scale effects change, TFPC1=EC+TC+SEC, AEIC=input market allocative efficiency change and TFPC=TFPC1+AEIC

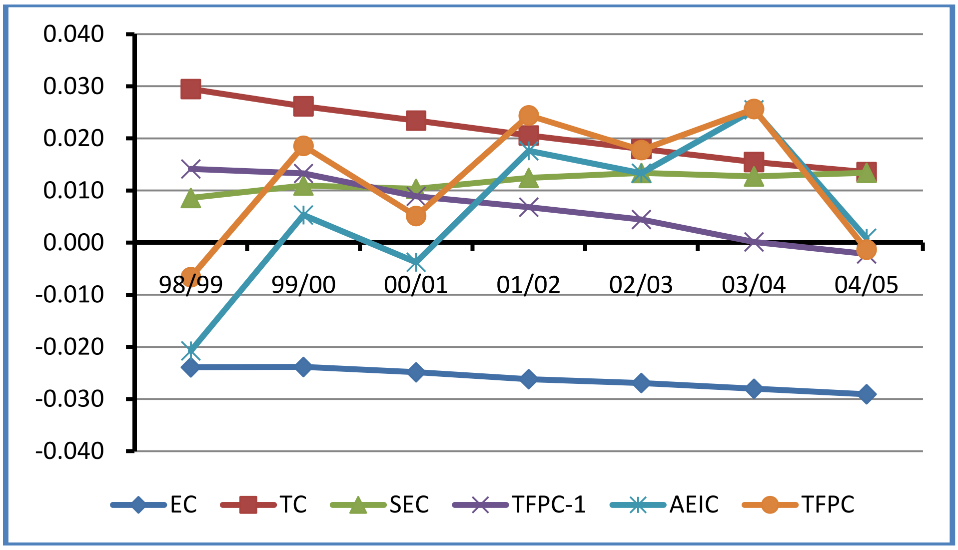

Figure 5.1: TFP change and its decomposition in Asian banking sector

Overall, Asian banking industries have experienced positive but not substantial productivity change for both sets of measurement of the TFP change during 1998-05. TFPC1 is about 0.6% and is mainly attributed to technical progress (2.1% on average) and economies of scale (1.2% on average) after offsetting the negative efficiency change (-2.6% on average). The negative sign of efficiency change indicates that cost efficiency has deteriorated throughout the sample periods. This trend is reflected by the value of η, estimated as -0.045 with a *t*-statistic of -14.958. In Battese and Coelli (1992) model, the statistically significant negative value of η suggests that cost inefficiency is increasing in an increasing rate, which leads to a decreasing trend of overall cost efficiency during the sample period. As discussed in chapter 4, the decline in efficiency may not necessarily imply that banks are performing worse over time as the estimation does not account for the quality of outputs (e.g. quality of loans). The attempts to remove large amount of NPLs after Asian financial crisis may result in a decreasing in net loan value as well as the measured cost efficiency. The positive 2.1% technical change suggests technical progress in Asian banking industry, indicating that banks benefit from using new technology in their business, such as the introduction of ATM, telephone banking and internet banking services. This result coincides with most productivity literature where technical progress is also found in the US and European banking sector. The positive scale effect change is consistent with the expectation as scale elasticity reported in Table 4.8 suggests that most Asian banks are operating with slight economies of scale. Therefore, economies of scale contribute around 1.2% per annum to TFP change. We also observe that TFPC1 is exhausted at the end of sample period due to the decreasing trend of technical progress. Although the positive scale effect change is stable the decreasing technical progress is not enough to offset the efficiency decline.

After adding-in the additional component of the allocative efficiency change, the TFP change improves further around 0.6% to reach 1.2% per year. As seen in Figure 5.1, the allocative efficiency is very volatile over the sample period and the whole pattern of TFP change has been highly influenced, implying that the allocative efficiency change is a very important source in explaining the TFP change. Once the information on input and output prices are available, it is necessary to calculate input and output price effects as additional sources of TFP change.

5.5.3 Country-specific TFP change and its decomposition

Results are presented in Table 5.3 and plotted in Figure 5.2 to 5.11. Without considering the contribution of allocative efficiency change in the TFP change index, the overall results of TFPC1 do show a modest productivity growth for Chinese, Hong Kong, Indian, Malaysian, Singaporean, Korean and Taiwanese banks (2%, 1.3%, 1.5%, 1.1%, 2.1%, 0.8% and 0.6% respectively). From analyzing the decomposition of TFPC1, the productivity growth in these banking sectors are mainly attributed to the net effects of technical progress and benefits from economies of scale after offsetting the downward trend of efficiency change. However, productivity deterioration has been brought for Indonesian, Philippine and Thai banks (-0.7%, -0.6% and -2.6% respectively) due to the net effects of decreasing efficiency change offsetting technical progress and economies of scale.

However, by including the allocative efficiency change in the calculation of TFP change, except for Korean banks (-2.2%), positive productivity growth has been found in all the other banking industries. However, the effect of allocative efficiency change on TFP growth is mixed. Clearly productivity growth for Chinese, Hong Kong, Indian, Malaysian banks has been strengthened with positive allocative efficiency change (increased by 1.1%, 0.4%, 1.3%, 0.1%). In addition, productivity deterioration for Indonesian, Philippine banks and Thai banks observed from TFPC1 has been turned around due to improvements in allocative efficiency of 0.9%, 1.9% and 3.9% respectively. However, for Singaporean and Taiwanese banks, their allocative efficiency change shows a reverse trend to their productivity growth for -0.5% and -0.3% respectively. Even worse for Korean banks, the substantial 3.0% allocative efficiency deterioration results in their productivity decline to -2.2%. This can partly be explained by the policies implemented during the reconstruction of Korean banking system after its systematic collapse in the Asian financial crisis. In order to reduce the substantial amount of NPLs, banks with a large volume of NPLs were forced to exit or to merge with healthy banks, leaving only 19 banks in 2005 compared with 33 banks at the end of 1997. Therefore, the profitability and performance of healthy banks is likely to have been weakened during the post-crisis period, reflected as a decline in the efficiency score. Also, after merging employees and large amount of fixed assets from bankrupt banks, new banks may operate well above their optimal mix of input quantities. At the same time, the attempts to ensure Korean banks satisfying the Basel capital adequacy ratio of 8% (with the fact that the capital adequacy ratio increased from 7% by the end of 1997 to 11.3% at the end of 2004), banks may be forced to buy a lot of equity capital which is proven to be far more costly than debt financing.

Although it is difficult to precisely explain the reasons behind the shift of productivity change between these Asian banking industries, some common scenarios can still be addressed. First of all, Asian banks have experienced an efficiency slump during 1998 to 2005. This creates the main deteriorating effect for TFP change. As explained before, in post-crisis recovery period, the attempts to reduce the large amount of NPLs, although inevitable and good in the long run will induce a large burden for banks from cost perspective, reflected as a decrease in cost efficiency. Second, there are technical progress and positive scale effect change in Asian banking sectors. These effects justify the current deregulation progress in most Asian banking industry. As found in many other studies (Alam, 2001; Berger *et al.*, 1992; Gilbert and Wilson, 1998; Kumbhakar and Lozano-Vivas, 2005; Mukherjee *et al.*, 2001; Tortosa-Ausina *et al.*, 2008), deregulation promotes technical progress by allowing banks to take advantage of technical advances. Moreover, deregulation allows banks to expand their loans and other business services with respect to the best interests of the shareholders and to

Table 5.3: Country-specific average TFP change and its decomposition

	Years	EC	TC	SEC	TFP-1	AEIC	TFP
China	98/99	-0.027	0.030	0.012	0.016	-0.006	0.010
	99/00	-0.026	0.028	0.030	0.032	0.012	0.044
	00/01	-0.026	0.026	0.028	0.029	0.005	0.034
	01/02	-0.026	0.023	0.034	0.031	0.057	0.088
	02/03	-0.027	0.021	0.027	0.021	0.005	0.026
	03/04	-0.027	0.019	0.013	0.004	0.007	0.012
	04/05	-0.029	0.016	0.021	0.008	-0.006	0.003
	98/05	-0.027	0.023	0.023	0.020	0.011	0.031
Hong Kong	98/99	-0.019	0.030	0.007	0.018	-0.023	-0.004
	99/00	-0.022	0.028	0.012	0.018	-0.007	0.011
	00/01	-0.023	0.025	0.004	0.005	-0.006	0.000
	01/02	-0.024	0.023	0.018	0.016	0.018	0.034
	02/03	-0.025	0.021	0.019	0.015	0.051	0.066
	03/04	-0.026	0.018	0.018	0.010	0.027	0.036
	04/05	-0.023	0.016	0.015	0.008	-0.032	-0.024
	98/05	-0.023	0.023	0.013	0.013	0.004	0.017
India	98/99	-0.009	0.026	0.010	0.028	0.049	0.077
	99/00	-0.012	0.022	0.013	0.023	0.000	0.023
	00/01	-0.012	0.019	0.010	0.017	-0.002	0.014
	01/02	-0.013	0.016	0.013	0.016	-0.008	0.009
	02/03	-0.013	0.012	0.009	0.007	0.008	0.015
	03/04	-0.014	0.010	0.015	0.011	0.029	0.039
	04/05	-0.015	0.008	0.010	0.003	0.013	0.016
	98/05	-0.012	0.016	0.011	0.015	0.013	0.028
Indonesia	98/99	-0.024	0.023	-0.001	-0.002	-0.011	-0.013
	99/00	-0.026	0.022	0.010	0.007	0.009	0.016
	00/01	-0.028	0.020	0.004	-0.004	-0.009	-0.013
	01/02	-0.029	0.017	0.003	-0.010	0.020	0.011
	02/03	-0.031	0.015	0.007	-0.010	0.014	0.004
	03/04	-0.032	0.013	0.008	-0.011	0.039	0.028
	04/05	-0.035	0.011	0.005	-0.019	-0.004	-0.023
	98/05	-0.029	0.017	0.005	-0.007	0.008	0.002
Malaysia	98/99	-0.016	0.027	0.015	0.027	-0.055	-0.029
	99/00	-0.018	0.025	0.007	0.014	0.004	0.018
	00/01	-0.020	0.023	0.016	0.019	-0.004	0.015
	01/02	-0.022	0.020	0.003	0.001	0.001	0.002
	02/03	-0.022	0.018	0.010	0.006	0.027	0.034
	03/04	-0.023	0.015	0.012	0.004	0.016	0.020
	04/05	-0.024	0.012	0.017	0.005	0.021	0.026
	98/05	-0.020	0.020	0.011	0.011	0.001	0.012

Table 5.3: Country-specific average TFP change and its decomposition (continued)

	Years	EC	TC	SEC	TFP-1	AEIC	TFP
Philippines	98/99	-0.025	0.026	0.008	0.009	0.012	0.020
	99/00	-0.026	0.023	0.002	-0.001	0.014	0.013
	00/01	-0.027	0.020	0.000	-0.007	0.011	0.003
	01/02	-0.028	0.017	0.010	-0.002	0.065	0.063
	02/03	-0.030	0.014	0.004	-0.012	0.000	-0.011
	03/04	-0.031	0.012	0.004	-0.016	0.014	-0.002
	04/05	-0.028	0.010	0.006	-0.013	0.018	0.005
	98/05	-0.028	0.017	0.005	-0.006	0.019	0.013
Singapore	98/99	-0.013	0.030	0.010	0.027	-0.047	-0.021
	99/00	-0.014	0.027	0.012	0.024	-0.002	0.022
	00/01	-0.014	0.024	0.022	0.031	-0.027	0.004
	01/02	-0.015	0.021	0.027	0.033	0.007	0.040
	02/03	-0.016	0.019	0.008	0.011	0.031	0.042
	03/04	-0.016	0.017	0.011	0.012	0.000	0.012
	04/05	-0.015	0.015	0.007	0.007	0.006	0.013
	98/05	-0.015	0.022	0.014	0.021	-0.005	0.016
Korea	98/99	-0.024	0.031	0.017	0.025	-0.126	-0.102
	99/00	-0.026	0.029	0.008	0.011	0.022	0.033
	00/01	-0.027	0.025	0.015	0.013	-0.048	-0.035
	01/02	-0.028	0.023	0.018	0.013	-0.031	-0.018
	02/03	-0.029	0.021	0.012	0.003	-0.015	-0.012
	03/04	-0.031	0.020	0.003	-0.008	-0.007	-0.015
	04/05	-0.033	0.020	0.014	0.001	-0.006	-0.004
	98/05	-0.028	0.024	0.012	0.008	-0.030	-0.022
Taiwan	98/99	-0.027	0.033	0.007	0.013	-0.013	0.001
	99/00	-0.028	0.031	0.009	0.011	-0.008	0.004
	00/01	-0.029	0.028	0.009	0.007	-0.014	-0.007
	01/02	-0.031	0.025	0.004	-0.001	-0.010	-0.012
	02/03	-0.032	0.023	0.020	0.011	0.004	0.015
	03/04	-0.033	0.019	0.021	0.007	0.021	0.028
	04/05	-0.034	0.016	0.014	-0.004	-0.003	-0.007
	98/05	-0.031	0.025	0.012	0.006	-0.003	0.003
Thailand	98/99	-0.049	0.035	0.008	-0.006	0.023	0.017
	99/00	-0.056	0.034	-0.003	-0.025	0.033	0.008
	00/01	-0.058	0.031	0.005	-0.022	0.038	0.016
	01/02	-0.062	0.028	0.002	-0.032	0.021	-0.012
	02/03	-0.066	0.025	0.005	-0.036	0.056	0.020
	03/04	-0.068	0.023	0.010	-0.034	0.117	0.082
	04/05	-0.070	0.022	0.020	-0.028	-0.018	-0.046
	98/05	-0.061	0.028	0.007	-0.026	0.039	0.012

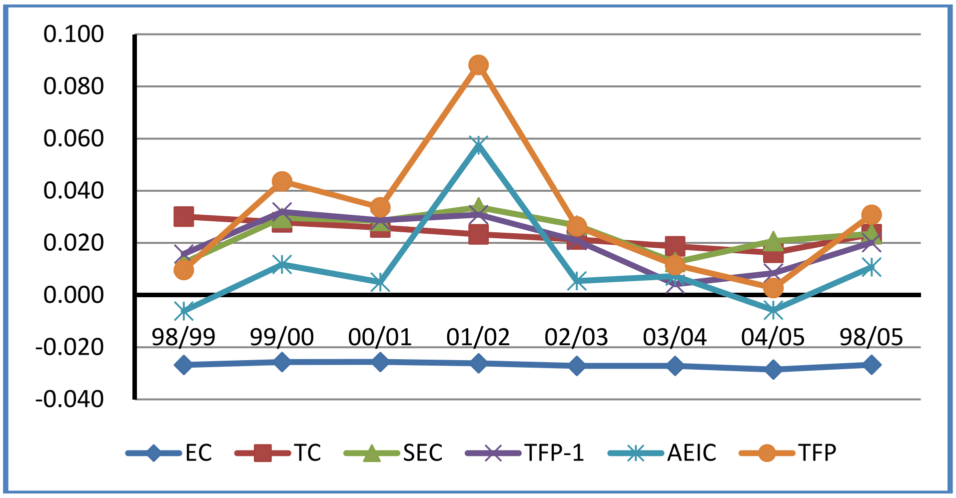

Figure 5.2: TFP change and its decomposition in Chinese banking sector

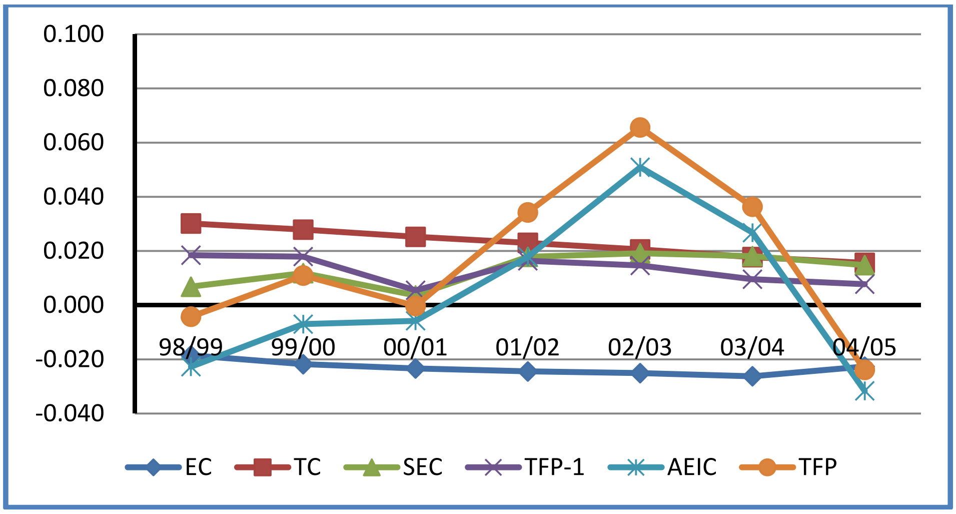

Figure 5.3: TFP change and its decomposition in Hong Kong's banking sector

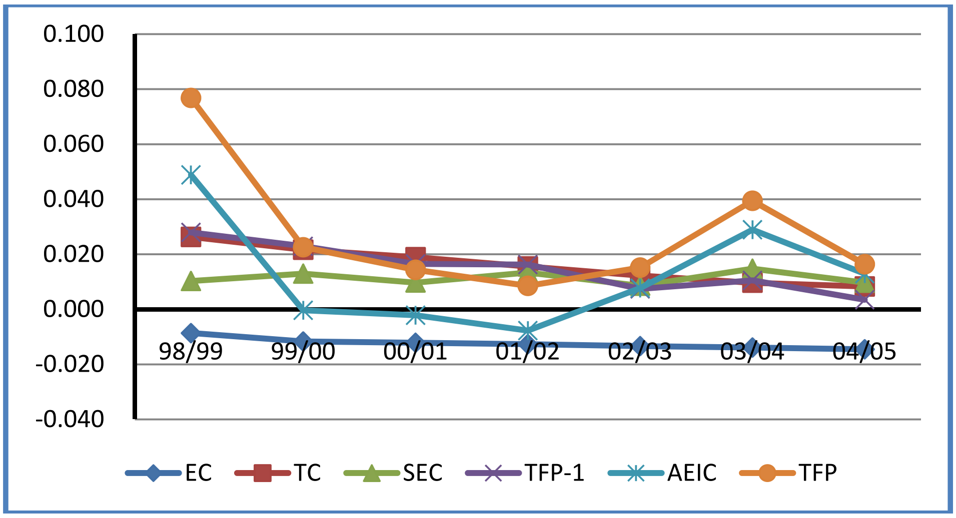

Figure 5.4: TFP change and its decomposition in Indian banking sector

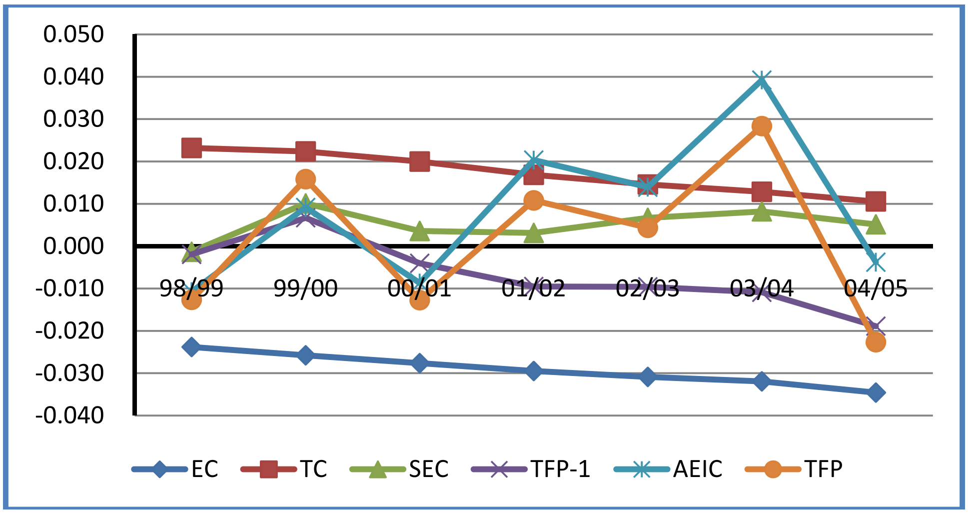

Figure 5.5: TFP change and its decomposition in Indonesian banking sector

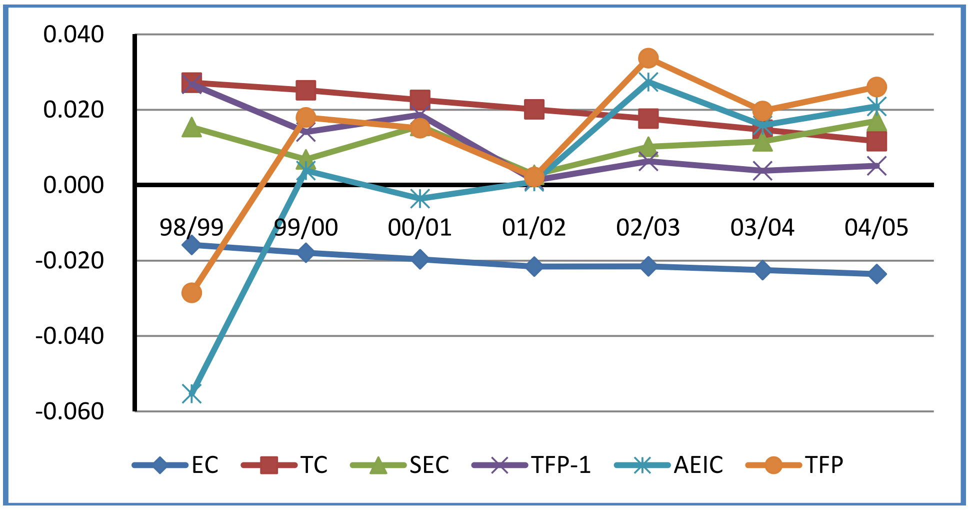

Figure 5.6: TFP change and its decomposition in Malaysian banking sector

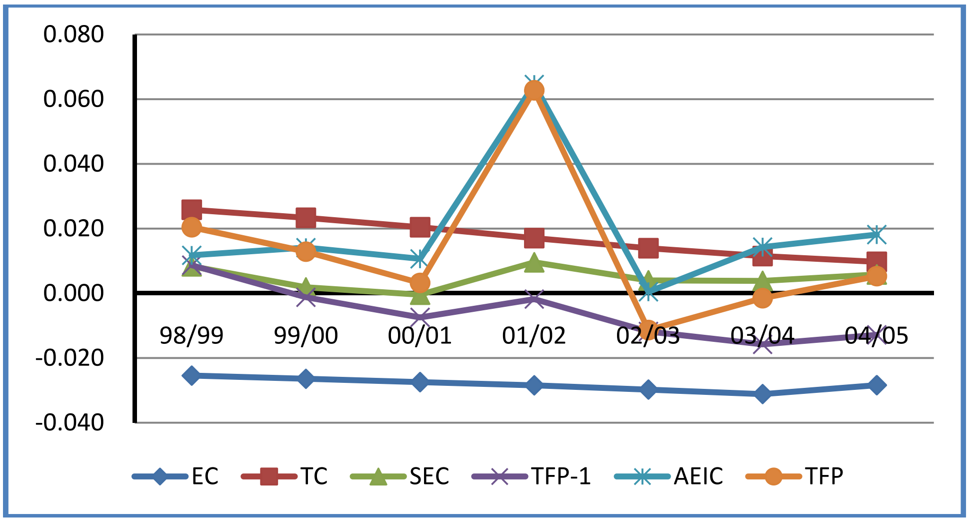

Figure 5.7: TFP change and its decomposition in Philippine banking sector

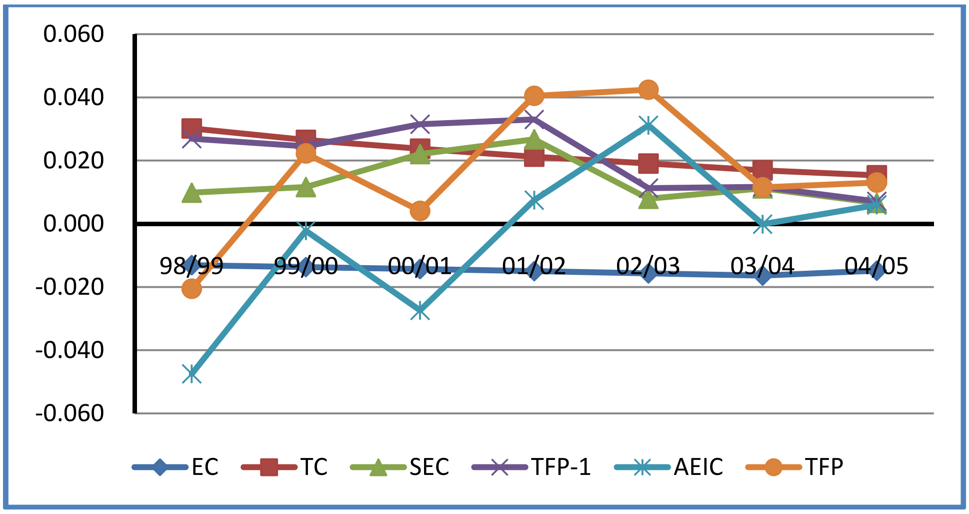

Figure 5.8: TFP change and its decomposition in Singaporean banking sector

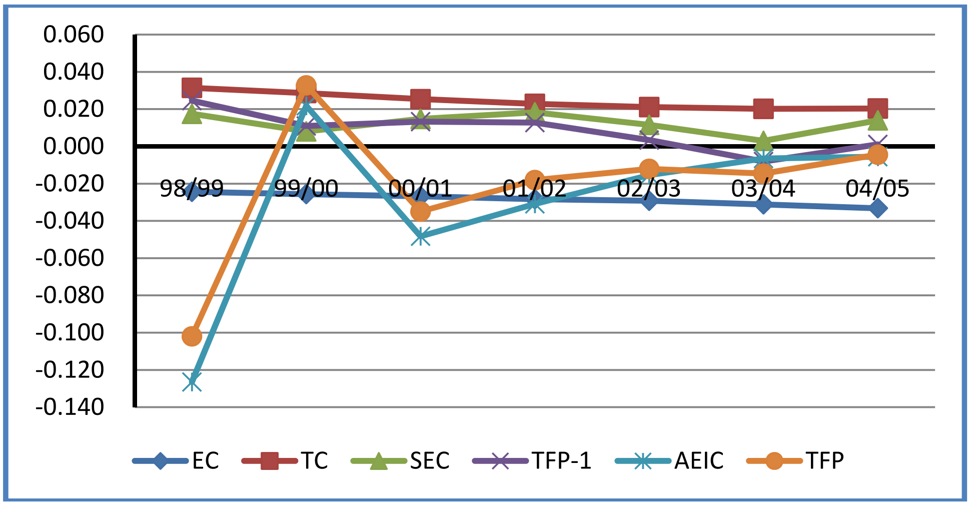

Figure 5.9: TFP change and its decomposition in Korean banking sector

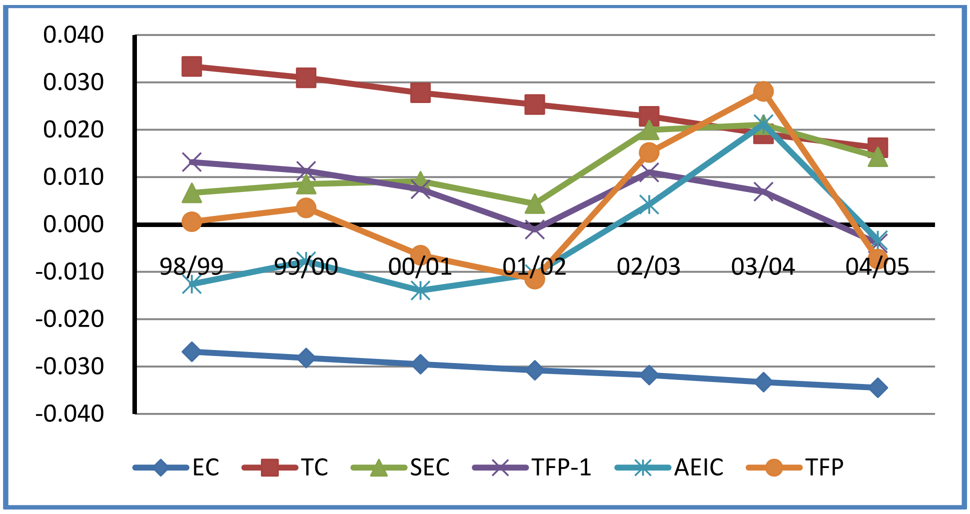

Figure 5.10: TFP change and its decomposition in Taiwanese banking sector

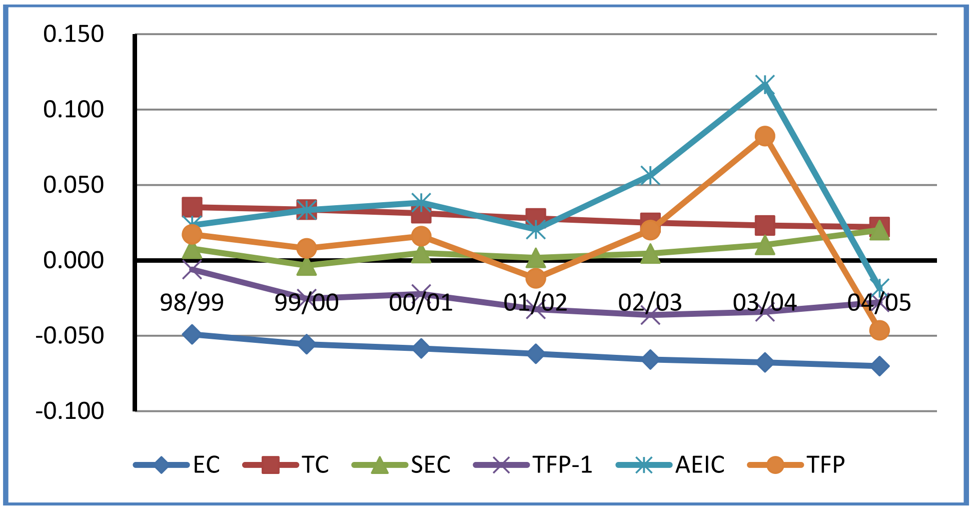

Figure 5.11: TFP change and its decomposition in Thai banking sector

pursue financial innovation. The positive scale effect change suggests that banks are operating with economies of scale, indicating that they benefit a lot from output expansion. Finally, Figure 5.2 to 5.11 provide a clear image that allocative efficiency change is a very important source to explain the productivity change. When it is excluded as in most previous productivity studies, TFP change index is quite smooth and show more or less the same trend for these Asian banks. However, the inclusion of allocative efficiency change presents a largely different picture as TFP change index is more volatile and largely influenced by the shape of allocative efficiency change.

5.6. Conclusion

The modern literature of measurement on productivity growth and its decomposition in banking sector has been dominated by studies in the US and European Union and by studies utilizing the non-parametric Malmquist productivity index, which usually decompose the TFP change into technical change, efficiency change and scale efficiency change. Limited attempts have tried to measure the productivity change in Asia but none of them use international data. Moreover, no banking studies have tried to measure the productivity change from the parametric stochastic cost frontier and to break down the TFP change with an additional source - allocative efficiency change. Therefore, this thesis is motivated to be among the first to address this issue by measuring the productivity change and its decomposition in major Asian banking industries during 1998 to 2005.

A parametric panel data stochastic cost frontier is estimated for calculating the TFP change. The new cost-based TFP index is derived from continuous time approach first proposed by Bauer (1990) and then converted to an index number approach in order to make the index applicable in the empirical study. The TFP change index is then decomposed into technical change, efficiency change, scale effect change and input mix and output mix allocative efficiency change.

Overall, Asian banking industries have experienced positive but not substantial productivity change during 1998-05, which is mainly attributed to net effect of technical progress and positive scale effect change after offsetting the decreasing trend of efficiency deterioration. However, productivity change is exhausted at the end of sample period due to the declining trend of technical change. Furthermore, the inclusion of allocative efficiency change influences the whole pattern of the TFP change, indicating that allocative efficiency change is a very important source to explain the productivity growth.

Chapter 6 Exogenous Influences on Inefficiency and Random Noise Components

6.1 Introduction

The previous two empirical chapters discuss the influences of cross-country exogenous environmental variables on the shape of production process, known as influences of heterogeneities on production technology. By relaxing the assumption of the same production technology usually shared in the efficiency literature, empirical results show that exogenous variables do influence the production process and neglecting them will put these exogenous influences as unobservable heterogeneities that inevitably are pushed into the inefficiency term therefore resulting in biased efficiency estimates. By modeling these exogenous influences as observable heterogeneities, one can obtain efficiency estimates that are net of environmental influences. As discussed in Chapter 3, influences of exogenous variables may not only be on the shape of production technology but also on inefficiency and random noise component in the stochastic frontier model. For example, exogenous variables can affect the mean of inefficiency, also known as heterogeneities in inefficiency in which researchers are interested to find explanation of the variations in efficiency in terms of exogenous variables. Besides the mean of inefficiency, exogenous influences on inefficiency can also be modeled as heteroscedasticity in the variance of inefficiency. Moreover, exogenous influences may be placed on the two-sided random noise component also as heteroscedasticity in the variance of the random noise term. Therefore, this chapter will discuss the issue of exogenous influences on inefficiency component and random noise.

In most of the efficiency literature that address exogenous influences on inefficiency, exogenous variables are modeled as contributors, or say determinants, of efficiency by employing a two-stage estimation. In the first stage, a stochastic frontier model and firm's efficiency level is estimated without considering exogenous variables. In the second stage, inefficiency will be regressed on these exogenous variables. However, such two-stage estimation provides biased results due to misspecification of the model; see discussions in section 2.1.2.6, Kumbhakar and Lovell (2000:264) and Wang and Schmidt (2002) for detailed discussion. The most famous and widely used one-stage approach is that of Kumbhakar *et al.* (1991), Huang and Liu (1994), and Battese and Coelli (1995) (KGMHLBC hereafter). In their approach, exogenous variables will influence the mean of the inefficiency term. Another set of one-stage approaches is that of Reifschneider and Stevenson (1991), Caudill and Ford (1993), Caudill *et al.* (1995) and Hadri (1999) (RSCFGH hereafter), in which exogenous variables will enter the variance of inefficiency. Wang (2002) combines these two by introducing a non-monotonic parameterization of exogenous variables in both mean and variance of inefficiency. On the other hand, exogenous variables could also influence the two-sided error term by entering its variance. This issue was first addressed by Hadri (1999) and further developed in Hadri *et al.* (2003a, 2003b). If exogenous variables do have an impact on the variance of the two-sided error term, ignoring it would provide biased estimates of parameters needed to calculating the efficiency and also parameters of the production process; see Kumbhakar and Lovell (2000: 116).

However, empirical studies combining all these three exogenous effects have rarely been seen in the modern literature of efficiency and productivity studies. The only study that can be found is Hadri *et al.* (2003a), in which the authors employ Battese and Coelli (1995) model with consideration of double heteroscedasticity in both error components to study the 101 mainly cereal farms in England. But so far, no studies appear in the banking efficiency and productivity literature, although studies with consideration of part of these three exogenous effects do exist. Therefore, I am motivated to analyze these three exogenous influences with an application to the Asian banking industries. To do so, a stochastic cost frontier is constructed and relative cost efficiency and TFP change will be estimated and calculated.

The rest of this chapter will be organized as follows. Section 6.2 will review the existing literature that addresses exogenous influences on inefficiency and random noise component. The general framework we intend to use and model specification will be discussed in section 6.3 and 6.4. Empirical results of model comparison, as well as cost efficiency and the TFP change will be presented in section 6.5. Finally, section 6.6 concludes.

6.2 Review of the literature

6.2.1 Modeling exogenous influences on inefficiency

The earlier prevailing approach modeling exogenous influences on inefficiency attempted to explain the variation of inefficiency (or efficiency) with the variation in exogenous variables. To do so, a two-stage procedure is employed. Assume a simple production frontier, in which a scalar output y is produced by a vector of inputs $\mathbf{x}$. Let $\mathbf{z}$ be a vector of exogenous variables which influence inefficiency. In the first stage, a stochastic production (or cost, or profit, or other) frontier is estimated (excluding exogenous variables) by the maximum likelihood estimation (MLE) under certain distributional and independence assumptions and the inefficiency estimates will be calculated by JLMS technique developed in Jondrow *et al.* (1982). Then in the second stage, the estimated inefficiencies (or efficiencies) will be regressed against a set of exogenous variables, usually in the form of linear regression.

$$
\begin{aligned}
&\ln y_{it} = \ln f(\mathbf{x}_{it};\boldsymbol{\beta}) + v_{it} - u_{it} \\
&E(u_{it} \mid v_{it} - u_{it}) = g(\mathbf{z}_{it};\boldsymbol{\beta}) + \varepsilon_{it} \\
&v_{it} \sim \text{iid } N(0,\sigma_v^2) \\
&u_{it} \sim \text{iid } N^+(0,\sigma_u^2)
\end{aligned}
\qquad [6.1]
$$

In this two-stage formulation [6.1], exogenous variables do not influence the structure of production frontier, but they do influence the efficiency with which producers approach the production frontier.

However, unfortunately there are two serious econometric problems associated with this two-stage formulation as summarized in Kumbhakar and Lovell (2000:264) and Wang and Schmidt (2002). First, the assumption of uncorrelateness of exogenous variables **z** and **x** must stand. If they are correlated, the maximum likelihood estimate of $(\boldsymbol{\beta}, \sigma_v^2, \sigma_u^2)$ is biased due to the omission of the relevant exogenous variables in the first stage estimation. Consequently the estimated efficiency is biased through the calculation involving those biased parameter estimates representing the production frontier. Therefore, it doesn't matter how the second stage regression is carried out since it will never provide the true determinants of the efficiency variation. The second problem lies in the contradiction in the treatment of inefficiencies. In the first stage, the inefficiencies are constructed as part of composed error term and are assumed to be independently identically distributed (e.g. half-normal, exponential, truncated-normal). But in the second stage, inefficiencies are modeled as a function of exogenous variables, which disobeys the presumed assumption in the first stage. Based on the Monte Carlo experiment, Wang and Schmidt (2002) suggest that this bias could be very severe and one-stage formulation that incorporates the stochastic production frontier and exogenous influences on inefficiency should be estimated simultaneously in the first step.

The single-stage stochastic frontier approaches that model exogenous influences on inefficiency can be categorized into two sets. The first set of models has the same objective to explain the variation of inefficiency with the variation of the exogenous variables. In those models, the one-sided error component representing inefficiency is assumed to follow the truncated normal distribution but the constant-mean property is relaxed to allow the mean of truncated normal distribution to be a function of the exogenous variables. This allows inefficiency, which depends on the mean of the truncated normal distribution, to depend on exogenous variables. Greene (2008) names this inefficiency effect as heterogeneity in the inefficiency term. The second set of approaches tries to relax the constant-variance property of truncated normal distribution by allowing the variance of inefficiency to be a function of exogenous variables. This allows the inefficiency, which depends on the variance of the truncated normal distribution, to depend on exogenous variables. This inefficiency effect is usually known as the problem of heteroscedasticity in inefficiency.

The first set of approaches that model the mean of the truncated normal distribution as a function of exogenous variables, consists of models such as KGMHLBC. These authors are inspired by the work from Deprins and Simar (1989a, b). In their paper they develop the first single-stage production frontier model that expresses the production frontier relationship as

$$\ln y_i = \ln f(\mathbf{x}_i; \boldsymbol{\beta}) - u_i \qquad [6.2]$$

$$E(u_i \mid \mathbf{z}_i) = \exp\{\boldsymbol{\gamma}'\mathbf{z}_i\} \qquad [6.3]$$

where $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are the vector of technology and environment parameters to be estimated and $u_i > 0$ presents inefficiency, and the exponential expression of relationship between inefficiency and exogenous variables ensures that $E(u_i \mid \mathbf{z}_i) > 0$. Then combining [6.2] and [6.3] and adding a random-noise error term yields the single-stage production frontier model

$$\ln y_i = \ln f(\mathbf{x}_i; \boldsymbol{\beta}) - \exp\{\boldsymbol{\gamma}'\mathbf{z}_i\} + \varepsilon_i \qquad [6.4]$$

However, the major difficulty with this approach is that it is based on the deterministic frontier model given in [6.2], which contains no systematic error component to capture the effects of random noise on production process. Therefore, a stochastic frontier model embedding this approach is more desirable and this leads to the development of KGMHLBC approaches.

Kumbhakar *et al.* (KGM for short) (1991) specify a stochastic production frontier model as

$$\ln y_i = \ln f(\mathbf{x}_i;\boldsymbol{\beta}) + v_i - u_i \qquad [6.5]$$

$$u_i = \boldsymbol{\gamma}'\mathbf{z}_i + \varepsilon_i \qquad [6.6]$$

which is different from Deprins and Simar formulation in two ways. First, a random noise term is introduced in the production process via the error component $v_i \sim \text{iid } N(0,\sigma_v^2)$ as shown in [6.5]. Second, a linear rather than an exponential expression of exogenous effects on inefficiency is introduced. Inserting [6.6] into [6.5] yields the single-stage production frontier model

$$\ln y_i = \ln f(\mathbf{x}_i;\boldsymbol{\beta}) + v_i - (\boldsymbol{\gamma}'\mathbf{z}_i + \varepsilon_i) \qquad [6.7]$$

To estimate the parameters $(\boldsymbol{\beta},\boldsymbol{\gamma})$ and the inefficiency, it is necessary to impose distributional assumptions on v_i and ε_i in order to derive the likelihood function. A simple approach is to assume that $v_i \sim \text{iid } N(0,\sigma_v^2)$ and $u_i \sim \text{iid } N^+(\gamma' z_i,\sigma_u^2)$, the one-sided error component representing the technical inefficiency has a truncated normal structure with mode dependent on z. Then the model parameters can be estimated by MLE with the log likelihood function of truncation normal mean of $\mu_i = \boldsymbol{\gamma}'\mathbf{z}_i$, written as

$$\text{lnL} = \text{constant} - \frac{N}{2}\ln(\sigma_v^2 + \sigma_u^2) - \sum_i \Phi\left(\frac{\gamma' z_i}{\sigma_u}\right) + \sum_i \Phi\left(\frac{\mu_i^*}{\sigma^*}\right) - \frac{1}{2}\sum_i \frac{(e_i + \gamma' z_i)^2}{\sigma_v^2 + \sigma_u^2} \qquad [6.8]$$

where $\mu_i^* = \dfrac{\sigma_v^2 \gamma' z_i - \sigma_u^2 e_i}{\sigma_v^2 + \sigma_u^2}, \sigma^* = \left(\dfrac{\sigma_v^2 \sigma_u^2}{\sigma_v^2 + \sigma_u^2}\right)^{1/2}$

and $e_i = \ln y_i - \ln f(\mathbf{x}_i;\boldsymbol{\beta})$ are the residuals obtained from estimating [6.7] and $\Phi(.)$ is the cumulative density function of a standard normal distribution. Then based on the estimated parameters $(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma_v^2,\sigma_u^2)$, technical inefficiency can be obtained by JLMS technique. These estimates are based on either the conditional mean or the conditional mode.

$$E(u_i \mid \varepsilon_i) = \mu_i^* + \sigma^* \frac{\phi(\mu_i^*/\sigma^*)}{\Phi(\mu_i^*/\sigma^*)} \qquad [6.9]$$

$$\text{Or} \quad M(u_i \mid \varepsilon_i) = \begin{cases} \mu_i^* & \text{if } \mu_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad [6.10]$$

where $\phi(.)$ is the probability density function of a standard normal distribution.

Huang and Liu (1994) specify a similar model to KGM except replacing $\boldsymbol{\gamma}'\mathbf{z}_i$ by $g(\mathbf{z}_i;\boldsymbol{\gamma})$. Their model is written as

$$\ln y_i = \ln f(\mathbf{x}_i;\boldsymbol{\beta}) + v_i - [g(\mathbf{z}_i;\boldsymbol{\gamma}) + \varepsilon_i] \qquad [6.11]$$

which can be estimated with the similar procedure in KGM. In Huang and Liu (1994), they include inputs $\mathbf{x}$ in the function of $g(\mathbf{z}_i, \mathbf{x}_i;\boldsymbol{\gamma})$ to allow the interactions between elements of $\mathbf{z}_i$ and $\mathbf{x}_i$. However, this model can be viewed as an improvement to KGM if $g(\mathbf{z}_i;\boldsymbol{\gamma})$ can be tested against $\boldsymbol{\gamma}'\mathbf{z}_i$.

Battese and Coelli (1995) have developed a model that is essentially the same as that of Huang and Liu (1994) but with two exceptions. First of all, it is constructed within a panel data context[15] and second, they do not include a set of inputs in their specification of $g(\mathbf{z}_i;\boldsymbol{\gamma})$. Their model consists of [6.5] and [6.6], written as

$$\ln y_{it} = \ln f(\mathbf{x}_{it};\boldsymbol{\beta}) + v_{it} - (\boldsymbol{\gamma}'\mathbf{z}_{it} + \varepsilon_{it}) \qquad [6.12]$$

The nonnegativity requirement $u_{it} = \boldsymbol{\gamma}'\mathbf{z}_{it} + \varepsilon_{it} \geq 0$ is modeled as $\varepsilon_{it} \sim \text{iid } N(0, \sigma^2_{\varepsilon it})$ with the distribution of ε_{it} being bounded below by the variable truncation $-\boldsymbol{\gamma}'\mathbf{z}_i$. Battese and Coelli note that this distributional assumption on ε_{it} is consistent with the distributional assumption on u_{it} that $u_{it} \sim \text{iid } N^+(\boldsymbol{\gamma}'\mathbf{z}_{it}, \sigma^2_{ut})$. Technical efficiency for the ith producer at

[15] Note however that this panel data setting is less general than it appears. Battese and Coelli (1995) maintain the assumption that individual error terms including the inefficiency component are identically and independently distributed, e.g. $u_{it} \sim \text{iid } N^+(\mathbf{r}'\mathbf{z}_{it}, \sigma^2_u)$. The independence assumption means that it is more correct to view this as a pooled model, so that this term "panel data context" can be assumed for a model in which a time-invariant structure is imposed as an additional assumption.

time t is defined by

$$TE_{it} = \exp(-u_{it}) = \exp(-\boldsymbol{\gamma}'\mathbf{z}_{it} - \varepsilon_{it}) \qquad [6.13]$$

The prediction of technical efficiency is based on its conditional expectation, given the model assumption. The formulation of this conditional technical efficiency estimates is provided in Kumbhakar and Lovell (2000: 271) as

$$E[\exp(-u_i) \mid (v_i - u_i)] = \left[\exp\left(-\mu_{*i} + \frac{1}{2}\sigma_*^2\right)\right] \cdot \left[\frac{\Phi[(\mu_{*i}/\sigma_*) - \sigma_*]}{\Phi(\mu_{*i}/\sigma_*)}\right] \qquad [6.14]$$

where $\mu_{*i} = \dfrac{\sigma_v^2(\boldsymbol{\gamma}'\mathbf{z}_i) - \sigma_u^2(\varepsilon_i)}{\sigma_v^2 + \sigma_u^2}$, $\quad \sigma_* = \left(\dfrac{\sigma_v^2\sigma_u^2}{\sigma_v^2 + \sigma_u^2}\right)^{1/2}$

In the other seemingly different vein of the literature, Reifschneider and Stevenson (1991), Caudill and Ford (1993), Caudill *et al.* (1995) and Hadri (1999) (RSCFGH hereafter) seek to address the exogenous influence on inefficiency as heteroscedasticity in the variance of inefficiency. The idea was first proposed but not implemented by Reifschneider and Stevenson (1991). Their formulation was written as follows:

$$\ln y_{it} = \ln f(\mathbf{x}_{it}; \boldsymbol{\beta}) + v_{it} - u_i \qquad [6.15]$$

with $v_i \sim$ iid $N(0, \sigma_v^2)$ as always, but with $u_{it} \sim$ iid $N^+(0, \sigma_{ui}^2)$ where $\sigma_{ui}^2 = \sigma_{uo}^2 + g(\mathbf{z}_i; \boldsymbol{\gamma})$ with $g(\mathbf{z}_i; \boldsymbol{\gamma}) \geq 0$.

Actually this formulation of $\sigma_{ui}^2 = \sigma_{uo}^2 + g(\mathbf{z}_i; \boldsymbol{\gamma})$ serves two purposes. Besides modeling exogenous influences on inefficiency, unmodeled heteroscedasticity can be corrected. As argued in Kumbhakar and Lovell (2000: 119), the unmodeled heteroscedasticity will lead to biased estimates of all parameters in the model, and hence to biased estimates of technical inefficiency of individual firm. Therefore, by modeling the variance of the inefficiency term as a function of exogenous variables, heteroscedasticity can be corrected.

This problem has also been proved by Caudill and Ford (1993) and Caudill *et al.* (1995) in which, the authors find that heteroscedasticity results in biased parameter estimates. Specifically, when the production function is estimated by MLE, heteroscedasticity leads to overestimation of the intercept and underestimation of the slope coefficients (The opposite should be found for cost frontier). In Caudill *et al.* (1995), the authors develop a cost frontier to model the heteroscedasticity. The cost function is specified as

$$TC_i = \mathbf{\beta}'\mathbf{x}_i + v_i + u_i \qquad [6.16]$$

where TC_i is total cost, $\mathbf{x}_i$ is the vector of explanatory variables including output quantities and input prices, $\mathbf{\beta}$ is the unknown set of parameters to be estimated and v_i is the two-sided error term that $v_i \sim \text{iid}\, N\left(0, \sigma_v^2\right)$. One-sided error term representing inefficiency u_i follows the distribution of $u_i \sim \text{iid}\, N^+\left(0, \sigma_{ui}^2\right)$, with $\sigma_{ui}^2 = \sigma \cdot \exp\left(\mathbf{\gamma}'\mathbf{z}_i\right)$. They also give a specific exponential functional form to the standard deviation of the two-sided error term so that $\sigma_v^2 = \exp(\theta)$.

The same approach is implemented by Hadri (1999) that reproduces the results obtained by Caudill *et al.* (1995). He suggests that the presence of heteroscedasticity should be corrected to avoid any misspecification error and the use of incorrect test. However, Hadri (1999) extends Caudill *et al.* (1995) by also considering the heteroscedasticity in the two-sided error term. More recently, Laureti (2008) adopts the same approach to model the exogenous influence in Human Capital Formulation in the Italian University System. The specific heteroscedastic frontier model enables one to consider the effect of students' individual characteristics and the influences of the resources and organization of the specific faculty on efficiency and his results suggest that the model specification is strongly supported by the data.

Unlike KGMHLBC models that propose parameterization that the mean of the truncated distribution depending on exogenous variables as a way to study the exogenous influences on inefficiency and RSCFGH models that address the inefficiency effects by a heteroscedastic stochastic frontier model with parameterization that the variance of truncated distribution of inefficiency is a function of exogenous variables, Wang (2002) considers a production approach with combination of these two kinds of approaches to study the exogenous influences on inefficiency. It is the first paper published that seeks to address the exogenous influences on inefficiency through the combined model. In his paper, Wang demonstrates the model's unique ability to accommodate the non-monotonic efficiency effects, which can be very important and useful in understanding the relationships between inefficiency and its exogenous determinants. Wang's (2002) general features of production frontier can be expressed as follows:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + (v_{it} - u_{it}) \quad [6.17]$$

$$v_{it} \sim \text{iid } N(0, \sigma_v^2) \quad [6.18]$$

$$u_{it} \sim \text{iid } N^+(\mu_{it}, \sigma_{it}^2) \quad [6.19]$$

$$\mu_{it} = \boldsymbol{\delta}'\mathbf{z}_{it} \quad [6.20]$$

$$\sigma_{it}^2 = \exp(\boldsymbol{\gamma}'\mathbf{z}_{it}) \quad [6.21]$$

This model encompasses KGMHLBC and RSCFGH as special cases. For KGMHLBC, this amounts to replacing $\mathbf{z}_{it}$ in [6.21] by a single constant of one; for RSCFGH, it is obtained by substituting zero for $\mathbf{z}_{it}$ in [6.20].

Alvarez *et al.* (2006) also study the stochastic frontier model in which observable characteristics of the firms affect their levels of technical inefficiency. Their special interests are on the model's satisfaction of scaling property, which suggests that the inefficiency term $u(\mathbf{z};\boldsymbol{\delta})$, as a function of firm characteristics **z**, can be written as a scaling function $h(\mathbf{z};\boldsymbol{\delta})$ times a random variable u^* that does not depend on **z**. This property implies that the changes in z affect the scale but not the shape of $u(\mathbf{z};\boldsymbol{\delta})$. To model and test this scaling property, Alvarez *et al.* (2006) adopts the similar model as Wang (2002) shown from [6.17] to [6.21] except for the specification of μ_{it}, in which they consider as exponential function of **z**,

$$\mu_{it} = \exp(\boldsymbol{\delta}'\mathbf{z}_{it}) \quad [6.22]$$

It also considers various special cases test the scaling property through relaxing the parameterization in [6.21] and [6.22]. Table 6.1 summarizes these special cases and scaling property is satisfied with the model specification of RSCFGH and scaled Stevenson model. In their empirical examples of Spanish savings banks and Indian farms, they strongly suggest that the problem of model selection really relies on the fitness of the data. In both of their empirical studies, the general model (Wang model but with $\mu_{it} = \exp(\boldsymbol{\delta}'\mathbf{z}_{it})$ is preferred to other special cases.

Table 6.1: Summary of stochastic frontier models incorporating the exogenous influences on inefficiency based on Alvarez *et al.* (2006)

Model	Restrictions	Specification for the inefficiency $u_{it} \sim N^{+}(\mu_{it}, \sigma_{it}^{2})$		Scaling Property
		Mean	Variances	
General Model (Wang)		$\mu_{it} = \mu \cdot \exp(\boldsymbol{\delta}'\mathbf{z}_{it})$	$\sigma_{it} = \sigma_u \cdot \exp(\boldsymbol{\gamma}'\mathbf{z}_{it})$	No
RSCFGH Model	$\mu = 0$	$\mu_{it} = 0$	$\sigma_{it} = \sigma_u \cdot \exp(\boldsymbol{\gamma}'\mathbf{z}_{it})$	Yes
KGMHLBC Model	$\gamma = 0$	$\mu_{it} = \mu \cdot \exp(\boldsymbol{\delta}'\mathbf{z}_{it})$	$\sigma_{it} = \sigma_u$	No
Scaled Stevenson Model	$\gamma = \delta$	$\mu_{it} = \mu \cdot \exp(\boldsymbol{\delta}'\mathbf{z}_{it})$	$\sigma_{it} = \sigma_u \cdot \exp(\boldsymbol{\gamma}'\mathbf{z}_{it})$	Yes
RSCFGH- μ Model	$\delta = 0$	$\mu_{it} = \mu$	$\sigma_{it} = \sigma_u \cdot \exp(\boldsymbol{\gamma}'\mathbf{z}_{it})$	No
Stevenson Model	$\delta = \gamma = 0$	$\mu_{it} = \mu$	$\sigma_{it} = \sigma_u$	N/A
Aigner *et al.* (1977) (ALS) Model	$\mu = \gamma = 0$	$\mu_{it} = 0$	$\sigma_{it} = \sigma_u$	N/A

6.2.2 Modeling exogenous influences on two-sided noise error term

This literature of exogenous influences on two-sided noise error term, also known as the problem of heteroscedasticity in two-sided error term, is relatively small and this issue is usually jointly tested with the exogenous influences on the inefficiency term. Hadri (1999) is the first paper to address this issue by extending the paper by Caudill *et al.* (1995) that study the heteroscedasticity in inefficiency. Hadri (1999) argues that two-sided noise error term may also be affected by the heteroscedasticity and that if the likely heteroscedasticity is ignored, it will lead to biased maximum likelihood estimators. To correct this potential misspecification error, Hadri (1999) extends the Caudill *et al.* (1995) approach by incorporating the heteroscedasticity in the two-sided error term as $\sigma_{vi}^2 = \exp(\boldsymbol{\theta}'\mathbf{z}_i)$ and the relevant log likelihood function will be different (see Hadri, 1999 for detailed log likelihood function). Applied with the same data as used by Caudill *et al.* (1995), the author suggests that the new specification is supported by the data and the firm specific inefficiency measures is extremely sensitive to the proposed correction for heteroscedasticity.

Hadri *et al.* (2003a, 2003b) estimate the efficiency of 102 mainly cereal farms in England for the year of 1982-1987 using Battese and Coelli (1995) model with the correction for double heteroscedasticity. The similar specification with incorporation of heteroscedasticity in the inefficiency term and two-sided noise error term is proposed as seen in Hadri (1999). Their results suggest that the specification of double heteroscedasticity is strongly supported by the data.

However, with the same specification of heteroscedasticity in both error terms, Wang (2002) indicates that the correction of heteroscedasticity in the two-sided error term is not favored by their annual data from 1975-1976 to 1984-1985 on farmers from the village of Aurepalle in India. Instead, the author concentrates on the model with only inefficiency term corrected for heteroscedasticity.

To summarize, the existing literature discussing the exogenous influences on the composed error terms has been dominated by studies focusing on the exogenous influences on inefficiency with the effort of either inserting the exogenous variables into the mean of the truncated normal distribution, or/and relaxing the constant-variance property of the truncated normal distribution to allow the variance of inefficiency to depend on exogenous variables. The limitation of studies on exogenous influences on the two-sided noise error term is justified by the fact that the correction of heteroscedasticity in the two-sided error term is strongly dependent on the data used. And in some study, it is reasonable to assume that heteroscedasticity is absent in the two-sided error term (Laureti, 2008).

6.2.3 Empirical application of modeling exogenous influences on composed error terms in the banking literature

A small volume of banking efficiency studies has addressed this issue. Caudill *et al.* (1995) and Hadri (1999) address the problem of heteroscedasticity in inefficiency using US bank cost data. However, their study is limited since the authors only account the presence of heteroscedasticity in the variance of truncated normal distribution of the inefficiency term. More recently, Alvarez *et al.* (2006) study the scaling property in models that inefficiency depends on the firm specific characteristics with an application to Spanish savings banks. However, their interests are mainly on testing the scaling property through model comparison. And since the presence of heteroscedasticity in the two-sided random error term is not relevant to the problem of scaling property, it is not discussed. Therefore, there is room for me to provide a systematic discussion of those models where both the one-sided inefficiency term and two-sided noise error term depend on the exogenous variables with the application to international banking data consisting of ten major Asian banking industries. After model comparison and statistic tests, a preferred model will be selected. Based on the parameter estimates from the preferred model, measurement of cost efficiency and the TFP change will be implemented.

6.3 The general framework

A stochastic cost frontier model combining features of potential exogenous influences on both inefficiency and random noise term, which are heterogeneities in inefficiency (KGMHLBC), heteroscedasticity in inefficiency (RSCFGH) and heteroscedasticity in the two-sided noise error term (Hadri 1999) is adopted and can be expressed as:

$$\ln C_{it} = \ln f\left(\mathbf{y}_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}, \boldsymbol{\gamma}\right) + v_{it} + u_{it} \quad [6.23]$$

$$v_{it} \sim \text{iid } N\left(0, \sigma_{vit}^2\right) \quad [6.24]$$

$$\sigma_{vit}^2 = \sigma_v^2 \cdot \exp\left(\varpi' \mathbf{z}_{it}\right) \quad [6.25]$$

$$u_{it} \sim \text{iid } N^+\left(\mu_{it}, \sigma_{uit}^2\right) \quad [6.26]$$

$$\mu_{it} = \mu \cdot \exp\left(\vartheta' \mathbf{z}_{it}\right) \quad [6.27]$$

$$\sigma_{uit}^2 = \sigma_u^2 \cdot \exp\left(\boldsymbol{\psi}' \mathbf{z}_{it}\right) \quad [6.28]$$

where C_{it} denotes the total cost for the ith producer at time t;

$\mathbf{y}_{it}$ and $\mathbf{w}_{it}$ is a vector of outputs and input prices for the ith producer at time t, respectively;

$\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the unknown vector of parameters to be estimated;

v_{it} is the two-sided noise error term with mean of zero and variances depending on the exogenous variables $\mathbf{z}_{it}$. In this study, following Hadri (1999), the specification of $\sigma_{vit}^2 = \sigma_v^2 \cdot \exp(\varpi'\mathbf{z}_{it})$ is used

u_{it} is the one-sided error term representing the inefficiency effect, which follows a non-negative truncated normal distribution which is associated with exogenous variables $\mathbf{z}_{it}$. Following Alvarez *et al.* (2006) specification, $\mu_{it} = \mu \cdot \exp(\vartheta'\mathbf{z}_{it})$ and $\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\boldsymbol{\psi}'\mathbf{z}_{it})$ are used;

ϖ , ϑ and ψ are the unknown vector of parameters to be estimated.

Table 6.2: A range of models to examine exogenous influences on inefficiency parameters and idiosyncratic error

Model	Restrictions	Specification for inefficiency $u_{it} \sim \text{iid } N^+(\mu_{it}, \sigma_{uit}^2)$		Specification for the random noise error term $v_{it} \sim \text{iid } N^+(0, \sigma_{vit}^2)$
		Mean	Variances	Variances
General Model		$\mu_{it} = \mu \cdot \exp(\vartheta'\mathbf{z}_{it})$	$\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\psi'\mathbf{z}_{it})$	$\sigma_{vit}^2 = \sigma_v^2 \cdot \exp(\varpi'\mathbf{z}_{it})$
Wang Model	$\varpi = 0$	$\mu_{it} = \mu \cdot \exp(\vartheta'\mathbf{z}_{it})$	$\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\psi'\mathbf{z}_{it})$	$\sigma_{vit}^2 = \sigma_v^2$
KGMHLBC Model	$\psi = \varpi = 0$	$\mu_{it} = \mu \cdot \exp(\vartheta'\mathbf{z}_{it})$	$\sigma_{uit}^2 = \sigma_u^2$	$\sigma_{vit}^2 = \sigma_v^2$
RSCFGH-μ Model	$\vartheta = \varpi = 0$	$\mu_{it} = \mu$	$\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\psi'\mathbf{z}_{it})$	$\sigma_{vit}^2 = \sigma_v^2$
RSCFGH Model	$\mu = \varpi = 0$	$\mu_{it} = 0$	$\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\psi'\mathbf{z}_{it})$	$\sigma_{vit}^2 = \sigma_v^2$
Homoscedastic Model	$\psi = \varpi = 0$	$\mu_{it} = \mu$	$\sigma_{uit}^2 = \sigma_u^2$	$\sigma_{vit}^2 = \sigma_v^2$

As summarized in Table 6.2, this general model encompasses six special cases:

1) Wang model adopted in Alvarez *et al.* (2006) that only considers the inefficiency effects with exogenous influences on the mean and variance of the inefficiency term. In this case, the restriction $\varpi = 0$ has to be specified from the general model.

The remaining five special cases of the general model all share the same assumption that exogenous influences only exist on the inefficiency term. The two-sided noise error term is assumed to be homoscedastic that follows the normal distribution with constant mean $\sigma_{vit}^2 = \sigma_v^2$. Since they all share this assumption, they can also be viewed as special cases of Wang model. With the similar specifications as appeared in Alvarez *et al.* (2006), these five special cases can be summarized as follows:

2) Scaled Stevenson model. It shares the same specification as Wang model that considers the exogenous influences on the mean and variance of inefficiency but it also satisfies the scaling property. In this case, the incorporated additional restriction $\vartheta = \psi$ guarantees the satisfaction of scaling property since the truncated normal distribution of the inefficiency term can then be written as $\exp(\mathbf{z}_{it}^{'}\vartheta) \cdot N^{+}(\mu, \sigma_u^2)$ in which $\exp(\mathbf{z}_{it}^{'}\vartheta)$ serves as the scaling factor $h(\mathbf{z}_{it}, \vartheta)$ that will not affect the shape of the inefficiency distribution determined by $N^{+}(\mu, \sigma_u^2)$.
3) KGMHLBC model. This is the special case of Wang model, in which only the mean of inefficiency will be influenced by exogenous variables. A further restriction of $\psi = 0$ has to be added into Wang model. Unfortunately, the scaling property is not satisfied since $\mathbf{z}_{it}$ will enter the mean which effectively affect the shape of distribution.
4) RSCFGH-μ Model. This can also be viewed as a special case of RSCFGH model, in which instead of half normal distribution in the inefficiency term, a truncated normal distribution with the mean of μ and variances of $\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\psi' \mathbf{z}_{it})$ is specified. Apparently, the scaling property is not satisfied.
5) RSCFGH model. This is the special case of Wang model, in which only the variance of the inefficiency term is affected by exogenous variables. An additional restriction $\mu = 0$ needs to be inserted in the specification of Wang model. RSCFGH model also

satisfies the scaling property because the parameterizations of $\sigma_{uit}^2 = \sigma_u^2 \cdot \exp(\psi' \mathbf{z}_{it})$ and $\mu_{it} = 0$ simply enable one to rewrite the distribution of the inefficiency as $\exp(\psi' \mathbf{z}_{it}) \cdot N^+(0, \sigma_u^2)$. Similar to scaled Stevenson model, $\exp(\psi' \mathbf{z}_{it})$ can be viewed as scaling factor $h(\mathbf{z}_{it}, \psi)$, while the shape of half normal distribution is not changed.

6) The homoscedastic model, in which exogenous influences on both inefficiency and the two-sided error term are not involved. Therefore, in this case, $\mu = \psi = 0$ needs to be added into the Wang model.

My main interest here is to find the best fitted model in these seven models. The test procedure can be described as follows. First, Wang model is against the general model that nested the heterogeneities in inefficiency and heteroscedasticity in both error components. Second, since the remaining five models (scaled Stevenson model, KGMHLBC model, RSCFGH-μ model, RSCFGH model and the homoscedastic model) are special cases of Wang model, instead of testing these five models against the general model, I test them against Wang model. If Wang model is preferred, I pick the right model based on the test results of Wang model against the general model. If one of the five models are preferred than Wang model, it will then be tested against the general model. In the test procedure, two commonly used statistical tests are adopted, which can be summarized as follows:

1) The Likelihood ratio test. The definition of the likelihood ratio test is provided in almost all Econometric textbooks. A brief principle of how likelihood ratio test is given here. Let $\ln L$ be the general notation for the logarithm of the likelihood function. Suppose that L_U is the maximized value of the likelihood when the model is estimated ignoring the restriction, and let L_R be the maximized value of the likelihood when the model is estimated with the restriction imposed. Then the test statistic is $LR = 2(\ln L_U - \ln L_R)$, the statistic is asymptotically distributed as a chi-square distribution χ_p^2, where p is the number of restrictions being tested.
2) The Wald test. The Wald test is a way of testing the significance of particular explanatory variables in a statistical model. If for a particular explanatory variable, or group of explanatory variables, the Wald test is significant then one would conclude that the parameters associated with these variables are not zero, so that the

variables should be included in the model. Let $\varsigma = 0$ to be the null hypothesis. Let $\hat{\varsigma}$ be the unrestricted estimate of ς and let $V(\hat{\varsigma})$ be the asymptotic variance-covariance matrix of this estimate. Then the test statistic is $\hat{\varsigma}'V(\hat{\varsigma})^{-1}\hat{\varsigma}$, and the statistic is also asymptotically distributed as a chi-square distribution χ_p^2, where p is the number of restrictions being tested.

Based on these two tests, the best fitted model is selected to measure the cost efficiency scores and the TFP change and its relevant sources. The cost efficiency estimates are obtained from the parameter estimates of the preferred stochastic cost model. In the meanwhile, the TFP change and its decomposition can be measured using the same approach constructed in last chapter.

6.4 Empirical application

Following the notation from [6.23] to [6.28], the model is specified as a translog stochastic cost function, same as the one used in Chapter 4 and 5, written as,

$$\begin{aligned}\ln C_{it} &= \alpha + \sum_{j=1}^{3}\beta_j \ln y_{jit} + \sum_{m=1}^{3}\delta_m \ln w_{mit} + \rho_i \ln E_{it} \\ &+ \frac{1}{2}\sum_{j=1}^{3}\sum_{k=1}^{3}\beta_{jk}\ln y_{jit}\ln y_{kit} + \frac{1}{2}\sum_{m=1}^{3}\sum_{n=1}^{3}\delta_{mn}\ln w_{mit}\ln w_{mit} + \sum_{j=1}^{3}\sum_{m=1}^{3}\gamma_{jm}\ln y_{jit}\ln w_{mit} \\ &+ \theta_1 t + \frac{1}{2}\theta_2 t^2 + \sum_{j=1}^{3}\xi_j \ln y_{jit} t + \sum_{m=1}^{3}\zeta_m \ln w_{mit} t + v_{it} + u_{it}\end{aligned}$$

[6.29]

The focus of this study is to examine the influences of exogenous variables on the inefficiency term and the two-sided noise error term. The set of exogenous variables consists of nine variables. Three of them (z1_GDP per capita, z2_inflation, z3_unemployment ratio) are included to capture the features of macroeconomic conditions and four of them (z4_banking concentration, z5_net interest margin, z6_average capital ratio and z7_intermediation ratio) are incorporated to indicate the regulatory environment and banking structure of different Asian banking industries. The other two variables are z8_bank size and z9_time trend. The inclusion of time trend variables will enable the inefficiency to be time-varying in a non-monotonic pattern. In this study, the following specifications for the parameterization of exogenous variables in both error components are adopted. By including a constant term in the mean of the truncation distribution of the inefficiency term, the parameterization of $\mu_{it} = \mu \cdot \exp\left(\mathbf{z}_{it}^{'}\vartheta\right)$ can be simplified as $\mu_{it} = \exp\left(\mathbf{z}_{it}^{'}\vartheta\right)$. Written in the full series of exogenous variables, the mean of the truncated distribution is

$$\mu_{it} = \exp\left(\vartheta_0 + \vartheta_1 z_1 + \vartheta_2 z_2 + \vartheta_3 z_3 + \vartheta_4 z_4 + \vartheta_5 z_5 + \vartheta_6 z_6 + \vartheta_7 z_7 + \vartheta_8 z_8 + \vartheta_9 z_9\right)$$

[6.30]

For the variance of the truncated distribution of inefficiency and variances of the two-sided error term, the parameterization of exogenous influences is specified as

$$\sigma_{uit}^2 = \sigma_u^2 \cdot \exp\left(\psi_1 z_1 + \psi_2 z_2 + \psi_3 z_3 + \psi_4 z_4 + \psi_5 z_5 + \psi_6 z_6 + \psi_7 z_7 + \psi_8 z_8\right)$$

[6.31]

$$\sigma_{vit}^2 = \sigma_v^2 \cdot \exp\left(\varpi_1 z_1 + \varpi_2 z_2 + \varpi_3 z_3 + \varpi_4 z_4 + \varpi_5 z_5 + \varpi_6 z_6 + \varpi_7 z_7 + \varpi_8 z_8\right)$$

[6.32]

Since the variable representing the time trend is omitted from the parameterization of variances of the truncated distribution of the inefficiency term, the issue of scaling property is not raised here because the scaling factor $\exp\left(\mathbf{z}_{it}^{'}\psi\right)$ and $\exp\left(\mathbf{z}_{it}^{'}\vartheta\right)$ is not identical[16]. Consequently, in the test procedure, scaled Stevenson model is dropped.

Homogeneity of degree one are imposed into [6.29]. Monotonicity and concavity conditions are checked as usual before obtaining any estimates of cost efficiency and components of the TFP change.

[16] The initial parameterization of the variance of inefficiency considers the influence of time trend. However, parameter coefficient associated with time trend is not significant and there is no improvement on the likelihood function as well.

6.5 Empirical results

6.5.1 Choice of model

Six stochastic cost frontier models, listed as the general model, the Wang model, KGMHLBC model, RSCFGH-μ model, RSCFGH model and the homoscedastic model are estimated using LIMDEP 9. To test whether there are exogenous influences on the two-sided noise error terms, I compare the general model with Wang model that only focuses on the exogenous influences on inefficiency term. Table 6.3 reports the estimated parameters for both models along with the likelihood ratio statistic and the Wald test statistic. The null hypothesis of homoscedasticity in the variance of the two-sided error term v_{it} is that the parameters for the exogenous variables in the variance of the two-sided error term should be jointly zero (H0: $\varpi_1 = \varpi_2 = ... = \varpi_8 = 0$, against the alternative that at least one parameter is different from zero). The likelihood ratio statistic is 167.62 and the Wald statistic is 324.31. Because the critical value of the chi-square distribution with ten degree of freedom is 23.21 at the 1% level, 18.31 at the 5% level and 15.99 at the 10% level, respectively, it clearly indicates the rejection of the null hypothesis. Therefore, this Asian banking data favours the general model in which exogenous variables not only influence the inefficiency term but also the two-sided error term. In other words, besides heterogeneities in inefficiency, there are also double heteroscedasticity in inefficiency and the random noise error term.

Next, the rest four models are tested against Wang model with the assumption that the two-sided error term is homoscedastic and only the inefficiency term is affected by the exogenous variables. As explained before, if Wang model is favored, the general model will be picked since Wang model is rejected based on the above test results. However, if any of the rest four models is favored than Wang model, it will be teseted against the general model. Table 6.4 presents the model estimates and statistical test value for this set of comparison.

First, KGMHLBC model against Wang model. If the test favors KGMHLBC model, it means that exogenous influences should only be placed in the mean of the truncated distribution of u_{it} as explanatory variables to explain the variation in the observed efficiency scores. In this case, the null hypothesis is that parameters of exogenous

Table 6.3: General nested model vs. Wang model

	General nested model		Wang model	
	Parameters	**t-statistics**	**Parameters**	**t-statistics**
Constant	0.3499***	24.555	0.3284***	20.462
Y1	0.5340***	53.382	0.5278***	49.948
Y2	0.4230***	36.993	0.4214***	34.728
Y3	0.0267***	3.892	0.0438***	5.818
W1	0.7987***	76.917	0.8040***	66.472
W2	0.1004***	12.053	0.1213***	12.433
lnE	-0.1034***	-12.152	-0.1069***	-11.950
Y11	0.0723***	27.855	0.0793***	32.675
Y12	-0.1537***	-21.490	-0.1754***	-27.898
Y13	0.0071*	1.678	0.0072**	2.262
Y22	0.1063***	23.926	0.1147***	32.718
Y23	-0.0396***	-8.748	-0.0392***	-10.670
Y33	0.0118***	6.419	0.0149***	8.448
W11	0.0077	1.398	0.0041	0.713
W12	0.0264***	2.773	0.0058	0.603
W22	-0.0061	-1.257	0.0047	0.970
Y1W1	-0.0482***	-7.187	-0.0523***	-8.250
Y1W2	0.0291***	5.020	0.0321***	5.886
Y2W1	0.0248***	3.125	0.0170**	2.359
Y2W2	-0.0396***	-5.518	-0.0289***	-4.433
Y3W1	0.0077	1.474	0.0143***	2.958
Y3W2	0.0081*	1.774	0.0023	0.534
T	-0.0393***	-7.861	-0.0307***	-5.276
SQRT	-0.0085***	-7.450	-0.0108***	-8.490
Y1T	-0.0110***	-3.915	-0.0144***	-5.399
Y2T	-0.0074**	-2.337	-0.0071***	-2.652
Y3T	0.0088***	5.517	0.0138***	9.499
W1T	-0.0069**	-1.961	-0.0115***	-3.443
W2T	0.0286***	9.464	0.0212***	7.000
Mean of truncated distribution of inefficiency u				
Constant	0.0333	1.292	-0.1419***	-2.918
Z2	-0.0866***	-3.492	-0.3769***	-8.631
Z3	-0.5980***	-2.745	-0.7166***	-3.586
Z4	0.3601***	8.990	0.0300	0.926
Z5	0.3916***	9.705	0.4607***	8.667
Z6	0.1254***	2.571	0.2265***	5.092
Z7	0.0508	1.298	0.1452***	4.049
Z8	0.3547***	7.052	0.7624***	10.812
lnTA	0.0351***	3.884	0.0111	1.350
T	0.0885***	6.083	0.1218***	9.555

Table 6.3: General nested model vs. Wang model (continued)

	General nested model		Wang model	
	Parameters	**t-statistics**	**Parameters**	**t-statistics**
Heteroscedasticity in the variance of truncated distribution u				
Z2	0.4061***	6.929	0.4125***	9.420
Z3	-6.6777***	-4.919	-5.0369***	-8.370
Z4	-0.5365***	-3.353	-0.6219***	-7.320
Z5	-1.4785***	-6.095	-0.6185***	-6.354
Z6	-1.5800***	-4.411	-0.6681***	-4.908
Z7	-0.9487***	-4.251	-0.3301***	-3.414
Z8	-2.3393***	-5.928	-1.3910***	-7.751
lnTA	-0.3624***	-11.110	-0.2874***	-13.562
Heteroscedasticity in the variance of two-sided error term v				
Z2	-0.1712***	-4.382		
Z3	1.2050***	4.994		
Z4	-0.6498***	-12.883		
Z5	0.4124***	5.653		
Z6	0.1145	1.363		
Z7	0.0574	0.901		
Z8	0.1821*	1.881		
lnTA	-0.1050***	-7.338		
L.L.F	685.3032		601.4946	
LR test			167.6172	
Wald test			324.3066	

Note: ***, **, * denote the rejection of the null at the 1%, 5% and 10% level.

variables shown in the variance of the truncated distribution of u_{it} should be jointly equal to zero (H0: $\psi_1 = \psi_2 = ... = \psi_8 = 0$). The results of the likelihood ratio test with a value of 428.26 and the Wald test with a value of 353.07 suggest that the null hypothesis is rejected and exogenous variables should be placed in the variance of the truncated distribution of u_{it}.

Second, RSCFGH-μ model against Wang model. It is to test whether the exogenous influences should only be incorporated in the variance of the truncated distribution of u_{it}. Therefore, the null hypothesis of the RSCFCFG-μ model against Wang model is that parameters of exogenous variables shown in the mean of the truncated distribution of u_{it} are jointly equal to zero (H0: $\vartheta_1 = \vartheta_2 = ... = \vartheta_9 = 0$). Based on the likelihood ratio of 465.94 and the Wald statistic of 348.65, the null hypothesis is rejected.

Moreover, to test whether there are no exogenous influences on the inefficiency term, represented by the homoscedastic model, the null hypothesis in this test is that parameters of the exogenous variables shown in the mean and variances of the truncated distribution of the inefficiency term u_{it} should be jointly equal to zero (H0: $\vartheta_2 = ... = \vartheta_9 = \psi_1 = ... = \psi_8 = 0$). The homoscedastic model is modeled as a random-effects model that inefficiency follows a truncated normal distribution with a constant mean and variance. But due the estimation problem in LIMDEP, I am only able to obtain the model estimates in the first stage that provides the parameter estimates in a cross-sectional context but unable to obtain the model estimates in the second stage that generates the parameters estimates in a panel data context. Therefore, our likelihood ratio test with a value of 763.89 is obtained as twice the difference of log likelihood function of Wang model and that of the homescedastic model in the cross-sectional context. However, estimated through the normal procedure, the Wald test has a value of 1533.84 indicating the rejection of null hypothesis. Therefore, the assumption of homoscedasticity in both error components does not stand.

To test the adequacy of RSCFGH model, one cannot directly test RSCFGH model against Wang model since restrictions on ϑ is not identifiable when $\mu = 0$, which in other words, one cannot simply impose a linear restriction in Wang model to test the adequacy of RSCFGH model. However, instead, one can test RSCFGH model against RSCFGH-μ model, in which the restriction of $\mu = 0$ will be tested. The null hypothesis of $\mu = 0$ is rejected by the likelihood ratio value of 52.23 and the Wald test statistics of 9348.46. Based on this result, RSCFGH model is outperformed by RSCFGH-μ model, which instead is outperformed by Wang model.

To summarize, the general model that incorporated exogenous influences on the mean of the truncated distribution of the inefficiency term but also accounted for the double heteroscedasticity in the inefficiency term and the two-sided error term appears to be the ‘best’ model. This finding coincides with Hadri *et al.* (2003a, 2003b) that justify the correction of heteroscedasticity problem in both error components. Therefore, to proceed, parameter estimates of the general model will be used to obtain the cost efficiency score and the TFP change and its sources.

Table 6.4: Wang model vs. its special cases

	Wang model		KGMLHBC model		RSCFGH-mu model		RSCFGH model		Homoscedastic model	
	Parameters	t-statistics	Parameters	t-statistics	Parameters	t-statistics	Parameters	t-statistics	Parameters	t-statistics
Constant	0.3284***	20.462	0.3527***	30.828	0.3672***	29.828	0.3254***	23.426	0.4379	0.149
Y1	0.5278***	49.948	0.5374***	49.194	0.5417***	48.526	0.5440***	47.536	0.5529***	41.646
Y2	0.4214***	34.728	0.4172***	34.318	0.4180***	32.915	0.4185***	31.875	0.3769***	25.749
Y3	0.0438***	5.818	0.0391***	4.944	0.0414***	5.197	0.0463***	5.754	0.0569***	6.433
W1	0.8040***	66.472	0.8124***	67.798	0.8104***	66.680	0.8066***	63.533	0.8052***	67.205
W2	0.1213***	12.433	0.1021***	11.537	0.1026***	11.395	0.1002***	10.799	0.0931***	10.373
lnE	-0.1069***	-11.950	-0.0923***	-9.539	-0.0900***	-8.903	-0.0852***	-8.181	-0.0872***	-8.358
Y11	0.0793***	32.675	0.0783***	29.716	0.0759***	27.280	0.0733***	25.374	0.0660***	20.345
Y12	-0.1754***	-27.898	-0.1727***	-24.021	-0.1672***	-21.460	-0.1545***	-19.582	-0.1348***	-15.456
Y13	0.0072**	2.262	-0.0002	-0.035	0.0012	0.263	-0.0052	-1.168	0.0085	1.576
Y22	0.1147***	32.718	0.1135***	28.298	0.1130***	25.573	0.1081***	23.951	0.1092***	17.753
Y23	-0.0392***	-10.670	-0.0296***	-7.058	-0.0340***	-7.638	-0.0364***	-7.962	-0.0655***	-9.891
Y33	0.0149***	8.448	0.0131***	5.862	0.0145***	6.408	0.0181***	8.429	0.0256***	9.781
W11	0.0041	0.713	0.0128**	2.109	0.0139**	2.130	0.0104	1.539	0.0124**	1.962
W12	0.0058	0.603	-0.0002	-0.020	-0.0013	-0.116	0.0036	0.311	-0.0092	-0.874
W22	0.0047	0.970	0.0063	1.129	0.0080	1.448	0.0077	1.369	0.0202***	3.667
Y1W1	-0.0523***	-8.250	-0.0504***	-7.854	-0.0518***	-7.511	-0.0549***	-7.442	-0.0553***	-7.752
Y1W2	0.0321***	5.886	0.0232***	3.881	0.0233***	3.590	0.0255***	3.695	0.0257***	3.846
Y2W1	0.0170**	2.359	0.0250***	3.125	0.0311***	3.571	0.0338***	3.620	0.0452***	5.201
Y2W2	-0.0289***	-4.433	-0.0289***	-3.885	-0.0335***	-4.180	-0.0359***	-4.205	-0.0434***	-5.307

Table 6.4: Wang model vs. its special cases (continued)

	Wang model		KGMLHBC		RSCFGH-mu		RSCFGH		Homoscedastic Model	
	Parameters	t-statistics	Parameters	t-statistics	Parameters	t-statistics	Parameters	t-statistics	Parameters	t-statistics
Y3W1	0.0143***	2.958	0.0099*	1.853	0.0087	1.613	0.0092*	1.653	0.0160***	2.692
Y3W2	0.0023	0.534	0.0093*	1.956	0.0108**	2.241	0.0094*	1.909	-0.0012	-0.232
T	-0.0307***	-5.276	-0.0066*	-1.748	-0.0019	-0.484	-0.0015	-0.367	-0.0032	-0.789
SQRT	-0.0108***	-8.490	-0.0024**	-2.068	-0.0028**	-2.315	-0.0032**	-2.507	-0.0029**	-2.245
Y1T	-0.0144***	-5.399	-0.0072**	-2.478	-0.0075**	-2.408	-0.0067**	-1.972	-0.0093**	-2.524
Y2T	-0.0071***	-2.652	-0.0105***	-3.317	-0.0110***	-3.247	-0.0110***	-2.982	-0.0078**	-1.975
Y3T	0.0138***	9.499	0.0130***	6.610	0.0128***	6.344	0.0121***	6.447	0.0100***	4.235
W1T	-0.0115***	-3.443	-0.0027	-0.795	-0.0062*	-1.712	-0.0081**	-2.086	-0.0106***	-2.678
W2T	0.0212***	7.000	0.0142***	4.428	0.0179***	5.175	0.0190***	5.197	0.0205***	6.041
Mean of truncated distribution of inefficiency u										
Constant	-0.1419***	-2.918	-17.1088***	-8.891	-15.5703***	-96.687				
Z2	-0.3769***	-8.631	0.0172	0.025						
Z3	-0.7166***	-3.586	-12.9161*	-1.820						
Z4	0.0300	0.926	1.1123	1.110						
Z5	0.4607***	8.667	0.7134	0.483						
Z6	0.2265***	5.092	0.2813	0.159						
Z7	0.1452***	4.049	-0.4259	-0.297						
Z8	0.7624***	10.812	4.6546**	2.331						
lnTA	0.0111	1.350	-2.2600***	-12.784						
T	0.1218***	9.555	1.4012***	5.656						

Table 6.4: Wang model vs. its special cases (continued)

	Wang model		KGMLHBC		RSCFGH-mu		RSCFGH		Homoscedastic Model	
	Parameters	**t-statistics**	**Parameters**	**t-statistics**	**Parameters**	**t-statistics**	**Parameters**	**t-statistics**	**Parameters**	**t-statistics**
Heteroscedasticity in the variance of truncated distribution u										
Z2	0.4125***	9.420			0.0357	1.438	0.0723**	2.200		
Z3	-5.0369***	-8.370			0.1969	0.694	0.2614	0.683		
Z4	-0.6219***	-7.320			0.0419	0.973	0.0002	0.003		
Z5	-0.6185***	-6.354			-0.0079	-0.122	0.0142	0.165		
Z6	-0.6681***	-4.908			0.0546	0.602	0.0454	0.376		
Z7	-0.3301***	-3.414			-0.0813	-1.235	-0.0842	-0.965		
Z8	-1.3910***	-7.751			0.1265	1.239	0.1237	0.897		
lnTA	-0.2874***	-13.562			-0.1221***	-7.674	-0.2232***	-9.428		
lnL	601.4946		387.3664		368.5265		342.4140		219.5484	
LR test			428.2564		465.9362		52.2250		763.8924	
Wald test			353.0673		348.6531		9348.4629		1533.8400	

Note: ***, **, * denote the rejection of the null at the 1%, 5% and 10% level, respectively.

Table 6.5: Monotonicity and concavity condition of the general model

Monotonicity Property	Elasticity	Parameters	Standard errors	Whole sample: % of sample points
at the sample mean	ey1	0.534	0.015	99.5
at the sample mean	ey2	0.423	0.010	98.7
at the sample mean	ey3	0.027	0.007	90.3
at the sample mean	ew1	0.799	0.010	100
at the sample mean	ew2	0.100	0.008	96.9
Scale Property	**Scale Elasticity**	**Standard errors**	**t-value**	**Whole sample: % of sample points with scale economies**
at the sample mean	E = 0.984	0.007	-2.346 Reject H_0: E=1	89.4
Concavity Property	**Objective function**	**Principal Minors**	**Values**	**Whole sample: % of sample where H(w) is negative definite**
at the sample mean	H(w)	First order	-0.145, -0.102	
		Second order	0.004	73.5

Note: Calculation of scale elasticities is based on equation 4.8 in p. 129.

6.5.2 Cost efficiency and TFP change measurement

The parameter estimates of the general model used to obtain the cost efficiency score and the TFP change are presented in the second column of Table 6.3. In order to obtain reliable efficiency and productivity estimates, it is necessary to see whether these parameters are truly estimated from the cost function. This can be done by checking the monotonicity, concavity condition of estimated cost function. Results are presented in Table 6.5. The monotonicity condition is satisfied at both the sample mean and the most sample points. The concavity condition is satisfied at the sample mean and 73.5% of sample. The model estimates also suggest the existence of economies of scale at the sample mean and 89.4% of sample. This finding of modest economies of scale is consistent with that found in previous two chapters where same set of exogenous variables influences are considered on the shape of production frontier. The elasticity of time trend variable that capture the shifts of the cost frontier has a value of -0.039. It is significant at the 1% level indicating about 3.9% technical progress per annum within the sample period. Also, all the interactive terms of time trend variable (with outputs and input prices) are significantly different from zero suggesting the non-neutral technical change in the Asian banking industries. This finding of significant neutral and non-neutral technical change[17] is also consistent with the results found in previous chapters.

Table 6.6: Average cost efficiency score from the general nested model and comparison with cost efficiency score from Battese and Coelli (1992) model with incorporation of cross-country environmental variables

	General nested model		Battese and Coelli	
	Efficiency score	**Ranks**	**Efficiency score**	**Ranks**
Whole Sample	0.5412	-	0.5897	-
Country-specific efficiency score				
China	0.5314	7	0.5860	5
Hong Kong	0.7154	2	0.6197	4
India	0.3565	10	0.7661	1
Indonesia	0.4742	9	0.5376	8
Malaysia	0.5624	5	0.6666	3
Philippines	0.5247	8	0.5475	6
Singapore	0.7090	3	0.7224	2
South Korea	0.6438	4	0.5397	7
Taiwan	0.7170	1	0.5121	9
Thailand	0.5385	6	0.2517	10

Overall cost efficiency score and country-specific cost efficiency score are reported in Table 6.6. The overall cost efficiency level is 0.5412 indicating around 46% of inputs are wasted in the production process. Compared with the efficiency estimates from Battese and Coelli (1992) model that incorporate the environmental variables in the model structure, not only the efficiency estimates show different levels but also do the rankings of the Asian banking industries. This difference partly comes from the differences of model specifications. Battese and Coelli (1992) model adopted and estimated in the previous two empirical chapters share the assumption that exogenous

[17] The distinction between neutral and non-neutral technical change is described in the classic paper by Hahn and Matthews (1964). In the context of the cost function, the model allows the impact of technical progress to represent a parallel downward shift in the cost function (neutral technical progress) or to represent a shift in the cost function that is either output saving or input saving depending on how the effect of the time variable interacts multiplicatively with the outputs or the input prices (non-neutral technical progress). In the results (Table 6.3 row 25-31) the statistical significance of all the interaction terms indicates non-neutral technical progress.

environmental variables influence the production technology and the shape of the cost frontier, while the general model adopted here measures the exogenous influences on both inefficiency and two-sided error terms. According to Coelli *et al.* (1999), the former is named as net efficiency in which the efficiency scores are net of environmental influences while the latter is named as gross efficiency in which efficiency estimates incorporate the exogenous influences. Moreover, differences in the efficiency estimates and ranking of the Asian banking industries can also be explained from the different specification in inefficiency. In Battese and Coelli (1992) model, inefficiency is parameterized as $u_{it} = \exp(-\eta(t-T)) \cdot u_i$. Therefore, although this parameterization allows inefficiency to be time-varying but this variation is restricted in a monotonic pattern (e.g. either increasing or decreasing in the sample period). However, in the general model, the parameterization of [6.30] enables the inefficiency term to change over time but in a non-monotonic pattern.

Table 6.7: Cost efficiency trend from the general nested model

	1998	1999	2000	2001	2002	2003	2004	2005	Average
Mean	0.6390	0.5083	0.4969	0.4901	0.5181	0.5512	0.5895	0.5788	0.5465
China	0.5423	0.4924	0.4966	0.4923	0.5093	0.5253	0.5607	0.5549	0.5217
Hong Kong	0.7172	0.7064	0.7503	0.7169	0.6968	0.6983	0.7116	0.7445	0.7178
India	-*	0.2822	0.2906	0.2875	0.3147	0.4204	0.4571	0.4247	0.3539
Indonesia	0.6665	0.4606	0.3057	0.3264	0.4527	0.4543	0.4862	0.6325	0.4731
Malaysia	0.5644	0.5392	0.5727	0.5467	0.5500	0.5495	0.5836	0.5694	0.5594
Philippines	0.5962	0.4741	0.4348	0.4612	0.4603	0.5402	0.6878	-	0.5221
Singapore	0.7222	0.7204	0.7317	0.6996	0.6965	0.6851	0.7062	0.7195	0.7102
Korea	0.5797	0.5754	0.6157	0.6266	0.6624	0.6636	0.7047	0.697	0.6406
Taiwan	0.7303	0.7285	0.7338	0.6983	0.6924	0.7093	0.7244	0.7356	0.7191
Thailand	-	0.4701	0.5095	0.4838	0.539	0.5656	0.5916	0.5931	0.5361

Note: * denotes that cost efficiency scores for India in 1998, Philippines in 2005 and Thailand in 1998 is not considered due to small observation of banks and lack of consistence to the whole efficiency trend.

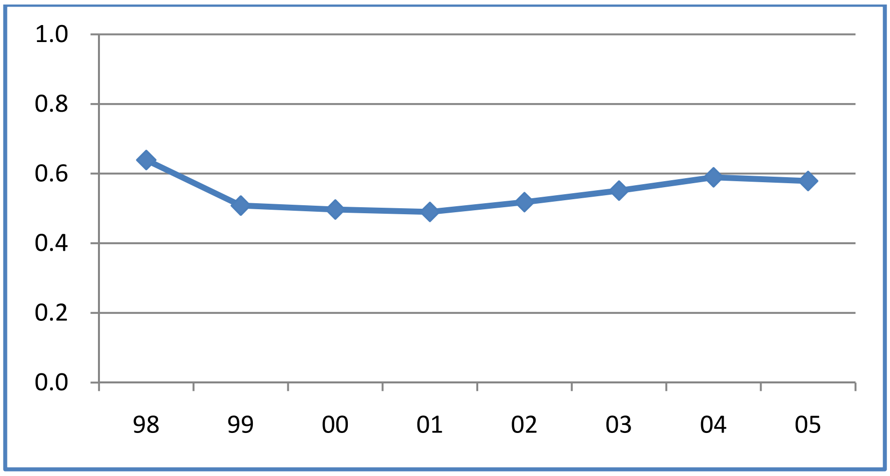

Figure 6.1: Overall Efficiency Trend

Table 6.7 [18] and Figure 6.1 to 6.11 reports the overall efficiency trend and country-specific efficiency trend from the general model. The overall efficiency trend is displayed in a non-monotonic pattern. It decreases from 0.6390 to 0.4901 from 1998 to 2001 and then increases to 0.5895 at the end of 2004 and slightly decreases in 2005. This efficiency trend actually captures the feature of the influence of the 1997 Asian financial crisis. All countries and regions in the sample data had been hit severely by the financial crisis. Heavily hit countries and regions had experienced the situations like currency depreciation, nation-wide economic meltdown and collapse of financial system. Less affected countries and regions like China, India, Singapore and Taiwan also suffered from a loss of demand and confidence throughout the region caused by the crisis contagion effect. Consequently commercial banks were also inevitably affected by the crisis since economic slump left commercial banks with large amount of non-performing loans (NPLs). As argued in chapter 4, the attempt to improve the loan quality will be reflected as a decrease in efficiency. These facts could explain the reason behind about 13% slump in the overall performance of these Asian banks in year 1998. In the post crisis period, most governments of these Asian countries and regions plan to stimulate the nation-wide economies through sound fiscal policies, to reconstruct the financial markets and to reform the banking system with the efforts to remove the

[18] Cost efficiency scores for India in 1998, Philippines in 2005 and Thailand in 1998 is not reported due to small observation of banks and lack of consistence to the whole efficiency trend.

massive amount of NPLs, to strengthen financial regulation and monitoring of banks' activities and to adopt the Basel accord capital adequacy ratio. In addition, the International Monetary Fund (IMF) also initiated a $40 billion rescue program to stabilize the currency. However, these effects did take time to make changes and improvement to these Asian economies. This explained why the efficiency level of these Asian banks decreased slightly in a decreasing rate at the beginning of the post crisis period 1999 to 2001 and then increased in an increasing rate from 2001 to 2004. Besides the factor of sound policy efforts made by these Asian governments and improvement of managerial performance in the individual bank, the improvement in the banking performance may also be reflected by the loan expansion from those Asian banks. From the market point of view, the reform and reconstruction of economies requires large amount of financial support. Industries, companies and even the government expect commercial banks to make out loans. From the banks' point of view, loans are still the main source of banks' income. The risk of loan expansion comes from the failure of repayment of loans that finally becomes the NPLs. However, considering the background of sound and booming economies, the default risk is relatively small. Besides, the central bank will always support the commercial banks as lender of last resort. Bank managers are motivated to make out more loans. With the given input level, activities of loan expansion will be reflected as an increase in cost efficiency. However, loan expansion should be carefully monitored since if loan expansion results in the massive excess of loans to deposits level, once loaners fail to repay their loan and the public panic turns to massive withdraw of their deposits, banks will have to face the risk of bankruptcy as seen in the most investment banks in the recent crisis of credit crunch.

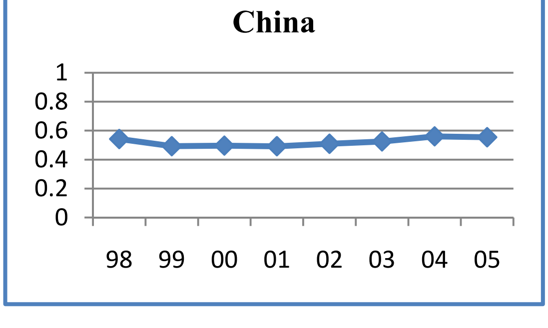

Figure 6.2: Efficiency trend for Chinese banking sector

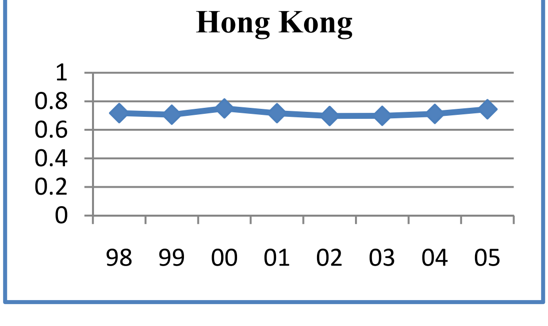

Figure 6.3: Efficiency trend for banking sector in Hong Kong

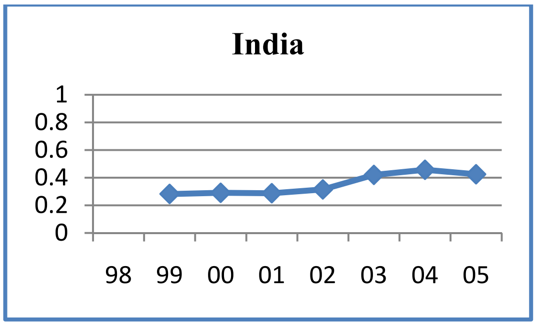

Figure 6.4: Efficiency trend for Indian banking sector

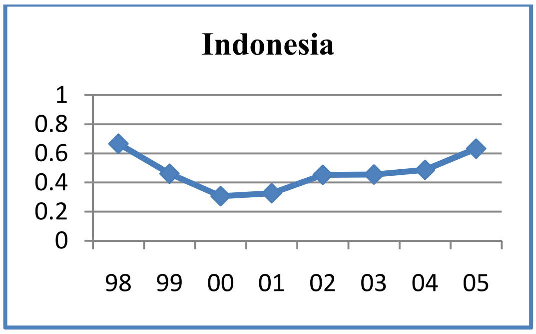

Figure 6.5: Efficiency trend for Indonesian banking sector

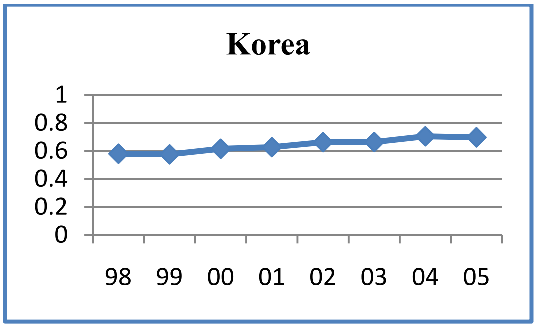

Figure 6.6: Efficiency trend for Korean banking sector

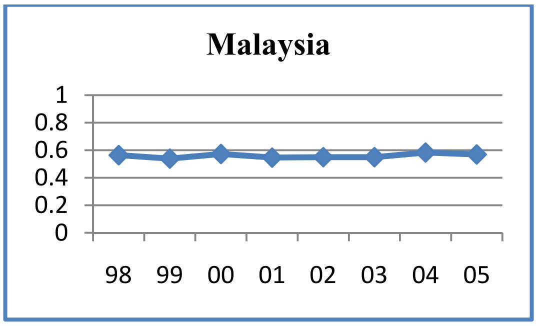

Figure 6.7: Efficiency trend for Malaysian banking sector

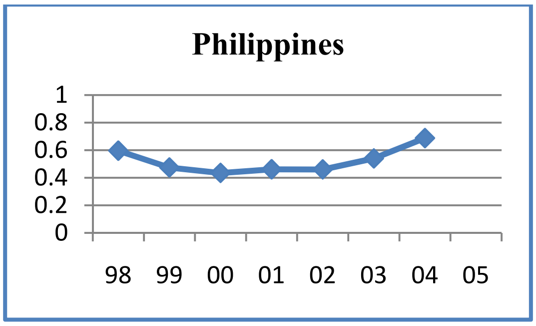

Figure 6.8: Efficiency trend for Philippine banking sector

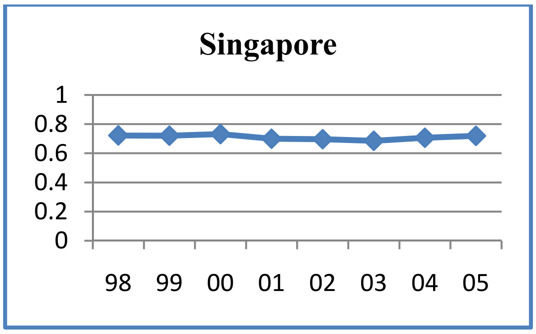

Figure 6.9: Efficiency trend for Singaporean banking sector

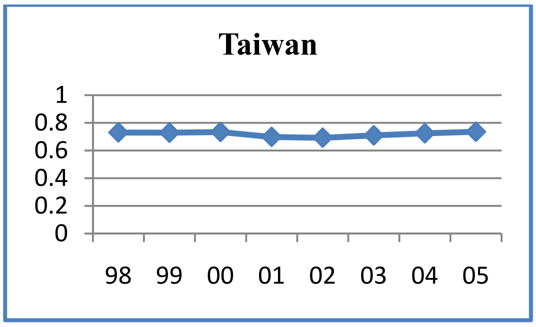

Figure 6.10: Efficiency trend for Taiwanese banking sector

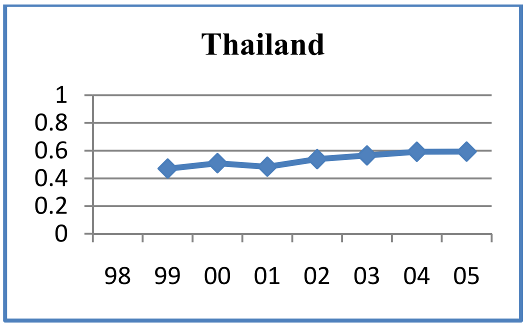

Figure 6.11: Efficiency trend for Thai banking sector

The country-specific efficiency trend shares the similar pictures as the overall efficiency trend is expected. An interesting observation is for Indonesian and Philippine banking sector. It seems that these two banking industries suffered more seriously than other Asian banking industries in the sample as their efficiency level experienced a heavily 20 and 12 percentage point slump from 1998 to 1999, respectively, compared to 0.5 to 3 percentage point slump in other Asian banking industries. As other Asian countries and regions started to recover from the financial crisis with 0.5 to 5 percentage point improvement in their banking performance from 1999 to 2000, Philippine banking sectors still shrank for about 3 percentage point. This may partly be explained by the fact that the Philippine economy heavily relies on the hot money and international investment. Due to the financial crisis, these foreign investments no doubt were withdrawn and this made the recovery of the Philippine economies more difficult and longer than other Asian economies. The efficiency slump for Indonesia is worse than Philippines as the Indonesian banking performance declined for about 16 percentage point from 1999 to 2000. Besides the same reason as happened to Philippines, the political turmoil was also attributed to this disastrous crisis since after 30 years in power, President Suharto was forced to step down in May 1998 in the wake of widespread rioting that followed sharp price increases caused by a drastic devaluation of the rupiah.

Table 6.8: Overall TFP change and its components

Period	EC	TC	SEC	AEIC	TFPC
1998-99	-0.0917	-0.0416	-0.0010	0.0472	-0.0871
1999-00	-0.0258	-0.0249	0.0020	0.0292	-0.0194
2000-01	-0.0435	-0.0080	0.0031	0.0191	-0.0293
2001-02	0.0393	0.0078	0.0048	0.0857	0.1376
2002-03	0.0923	0.0248	0.0043	0.0626	0.1841
2003-04	0.0886	0.0415	0.0051	0.0569	0.1920
2004-05	0.0175	0.0634	0.0054	-0.0116	0.0747
1998-05	0.0110	0.0090	0.0034	0.0413	0.0646

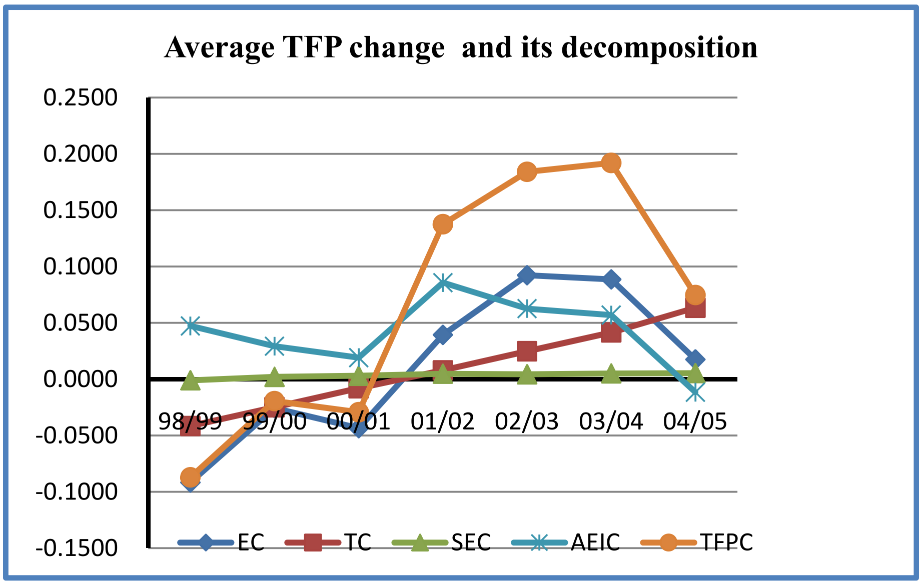

Figure 6.12: Overall TFP change and its decomposition in Asian banking sector

The measurement of the TFP change and its decomposition is carried out using the same method as in chapter 5. Table 6.8 and Figure 6.12 present the overall TFP change and its sources in these Asian banking industries over the sample period. The overall TFP change is around 6.5% per annum during the sample period. The overall TFP change can be further decomposed to four sources as cost efficiency change, technological change, change in scale effect and input mix allocative efficiency change. This about 6.5% overall TFP change comes from 1.1% cost efficiency change, 0.9% technological change, 0.3% scale effect change and 4.1% input mix allocative efficiency change. As seen in Figure 6.12, the overall TFP change follows the similar pattern as the cost efficiency change. In the other word, the shape of the TFP change is determined by cost efficiency change. While the scale of the TFP change depends on the sum of factors such as technological change, scale effect change and input mix allocative efficiency change. In a similar finding as addressed in the last chapter, input mix allocative efficiency change is the main contributor to the scale of the TFP change as it contributes about 4% to the overall 6% TFP change. Technological change shows a non-monotonic trend over the sample period with technological depress from 1998 to 2001 for the reason of financial crisis and then technological progress afterwards thanks to the recovery of Asian economies. Scale effect change shares the same increasing trend as technological change indicating that due to slight economies of scale found in the efficiency results, commercial banks benefit more and more from expanding their business and bank assets, mostly from the loan expansion.

Table 6.9[19] and Figure 6.13 to Figure 6.22 report the country-specific TFP change and its sources. Similar to the overall TFP change, country-specific TFP change shares the same pattern as cost efficiency change and its scale is determined by the joint effect of technological change, scale effect change and input mix allocative efficiency change. Different from the overall TFP change, the main contributors to the country-specific TFP change is different. For example, China has experienced the overall 9.6% productivity growth over the sample period and it mainly comes from the technological progress for about 5.2%. Malaysia, Philippines, Singapore, Taiwan and Thailand benefit from respective 4.3%, 7.2%, 2.2%, 3.9% and 13.5% input mix allocative efficiency change along with their productivity growth of 5.9%, 4.4%, 3.7%, 5.2% and 8.9% individually. For India, Hong Kong and South Korea, their respective overall 16.3%, 8.8% and 3.1% productivity growth is mainly from 13.2%[20], 5% and 2.9% inprovement in their cost efficiency performance, respectively. Although served as main contributor to the TFP change, cost efficiency change influences in the opposite direction for Indonesian banks as their 0.2% productivity growth are resulted from the largely offset of 6% cost efficiency decline to 4% increase in input mix allocative efficiency change. Its overall efficiency descent mainly comes from the massive slump of its efficiency from 1998 to 2000 for the reasons. Another interesting finding is that countries like China, India and Singapore that were less affected by the 1997 Asian financial crisis

---

[19] For Philippines, its TFP change is measured for the period of 1998 to 2004. This is due to small number of observations in 2005 (i.e. 23 observations for 2004 but 3 for 2005).

[20] Indian banks show a large volatility in cost efficiency change with a substantial improvement in cost efficiency from 2002 to 2003. Compared to the TFP change results in chapter 5, where cost efficiency changes are relative flat throughout the sample period, this large volatility may be attributed to the exogenous effects in the inefficiency term. However, this issue has never been discussed in the literature. Its explanation and driven sources are worth studying in the future research.

Table 6.9: Country-specific TFP change and its decomposition

	Years	EC	TC	SEC	AEIC	TFPC
China	98/99	-0.0283	0.0158	-0.0041	0.0294	0.0128
	99/00	0.0140	0.0221	-0.0065	0.0455	0.0751
	00/01	-0.0087	0.0351	-0.0034	0.0361	0.0592
	01/02	0.0359	0.0476	-0.0004	0.1027	0.1858
	02/03	0.0227	0.0651	0.0005	0.0052	0.0934
	03/04	0.0603	0.0803	0.0016	0.0046	0.1468
	04/05	-0.0071	0.0975	0.0043	0.0009	0.0955
	98/05	0.0127	0.0519	-0.0012	0.0321	0.0955
Hong Kong	98/99	-0.0201	-0.0637	0.0021	0.0387	-0.0430
	99/00	0.0561	-0.0521	0.0050	0.0058	0.0147
	00/01	-0.0110	0.0017	0.0040	-0.0015	-0.0068
	01/02	-0.0172	0.0194	0.0074	0.0031	0.0125
	02/03	0.2967	0.0354	0.0050	0.0323	0.3694
	03/04	0.0946	0.0528	0.0081	0.0508	0.2062
	04/05	-0.0525	0.0725	0.0043	0.0372	0.0615
	98/05	0.0495	0.0094	0.0051	0.0238	0.0878
India	98/99	0.0134	-0.0724	0.0066	0.0547	0.0023
	99/00	0.0186	-0.0152	0.0055	0.0037	0.0126
	00/01	0.0218	-0.0162	-0.0008	0.0014	0.0062
	01/02	0.2720	0.0006	-0.0007	0.0057	0.2776
	02/03	0.5283	0.0160	-0.0001	0.0382	0.5824
	03/04	0.0634	0.0338	-0.0007	0.0605	0.1570
	04/05	0.0049	0.0561	-0.0014	0.0403	0.0999
	98/05	0.1318	0.0004	0.0012	0.0292	0.1626
Indonesia	98/99	-0.5207	-0.0532	-0.0105	0.1300	-0.4544
	99/00	-0.4255	-0.0240	0.0095	0.0600	-0.3799
	00/01	-0.2146	-0.0030	0.0079	-0.0298	-0.2395
	01/02	0.2391	0.0191	0.0057	0.0312	0.2951
	02/03	0.0624	0.0384	0.0066	0.0604	0.1677
	03/04	0.1704	0.0508	0.0047	0.0842	0.3102
	04/05	0.2702	0.0715	0.0041	-0.0323	0.3134
	98/05	-0.0598	0.0142	0.0040	0.0434	0.0018
Malaysia	98/99	-0.0284	-0.0466	0.0000	0.1060	0.0310
	99/00	0.0587	-0.0236	0.0004	0.0367	0.0723
	00/01	-0.0466	-0.0086	0.0029	0.0071	-0.0452
	01/02	0.0095	0.0086	0.0005	0.0678	0.0864
	02/03	-0.0051	0.0291	0.0043	0.0532	0.0814
	03/04	0.0605	0.0492	0.0024	0.0399	0.1520
	04/05	-0.0313	0.0713	0.0046	-0.0121	0.0325
	98/05	0.0025	0.0113	0.0021	0.0427	0.0586

Table 6.9: Country-specific TFP change and its decomposition (continued)

	Years	EC	TC	SEC	AEIC	TFPC
Philippines	98/99	-0.1852	-0.0549	0.0023	0.0576	-0.1801
	99/00	-0.1184	-0.0352	0.0012	0.0117	-0.1408
	00/01	0.0433	-0.0171	-0.0005	0.0067	0.0324
	01/02	0.0262	-0.0013	0.0077	0.1680	0.2006
	02/03	0.1622	0.0169	0.0026	0.0053	0.1869
	03/04	0.2741	0.0374	0.0049	0.0146	0.3310
	04/05	-	-	-	-	-
	98/04	0.0337	-0.0090	0.0030	0.0440	0.0717
Singapore	98/99	-0.0025	-0.0277	-0.0002	0.0276	-0.0028
	99/00	0.0155	-0.0143	-0.0001	0.0000	0.0011
	00/01	-0.0448	0.0025	-0.0001	0.0447	0.0023
	01/02	-0.0045	0.0130	0.0023	0.0877	0.0985
	02/03	-0.0165	0.0255	0.0002	0.1048	0.1140
	03/04	0.0303	0.0460	0.0013	-0.0141	0.0634
	04/05	0.0096	0.0690	0.0017	-0.0965	-0.0162
	98/05	-0.0018	0.0163	0.0007	0.0220	0.0372
Korea	98/99	0.0152	-0.0749	0.0046	-0.0634	-0.1185
	99/00	0.0740	-0.0450	0.0018	0.0543	0.0852
	00/01	0.0124	-0.0256	0.0044	-0.0227	-0.0315
	01/02	0.0539	-0.0116	0.0058	0.0275	0.0756
	02/03	0.0016	0.0085	0.0046	0.0298	0.0445
	03/04	0.0558	0.0281	0.0008	0.0134	0.0980
	04/05	-0.0073	0.0494	0.0048	0.0151	0.0619
	98/05	0.0294	-0.0102	0.0038	0.0077	0.0307
Taiwan	98/99	-0.0069	-0.0261	0.0004	0.0141	-0.0185
	99/00	0.0087	-0.0121	0.0012	0.0055	0.0034
	00/01	-0.0510	0.0017	0.0032	0.0177	-0.0284
	01/02	-0.0099	0.0108	0.0011	0.0858	0.0878
	02/03	0.0229	0.0180	0.0061	0.1170	0.1640
	03/04	0.0210	0.0281	0.0068	0.0706	0.1266
	04/05	0.0169	0.0456	0.0060	-0.0373	0.0312
	98/05	0.0002	0.0094	0.0035	0.0390	0.0523
Thailand	98/99	0.0147	-0.1412	-0.0221	0.2768	0.1281
	99/00	0.0799	-0.1124	-0.0047	0.1380	0.1009
	00/01	-0.0506	-0.0990	0.0064	0.1087	-0.0345
	01/02	0.1156	-0.0818	0.0089	0.0646	0.1073
	02/03	0.0561	-0.0694	0.0032	0.1514	0.1414
	03/04	0.0334	-0.0541	0.0085	0.2800	0.2678
	04/05	0.0079	-0.0335	0.0175	-0.0781	-0.0861
	98/05	0.0367	-0.0845	0.0025	0.1345	0.0893

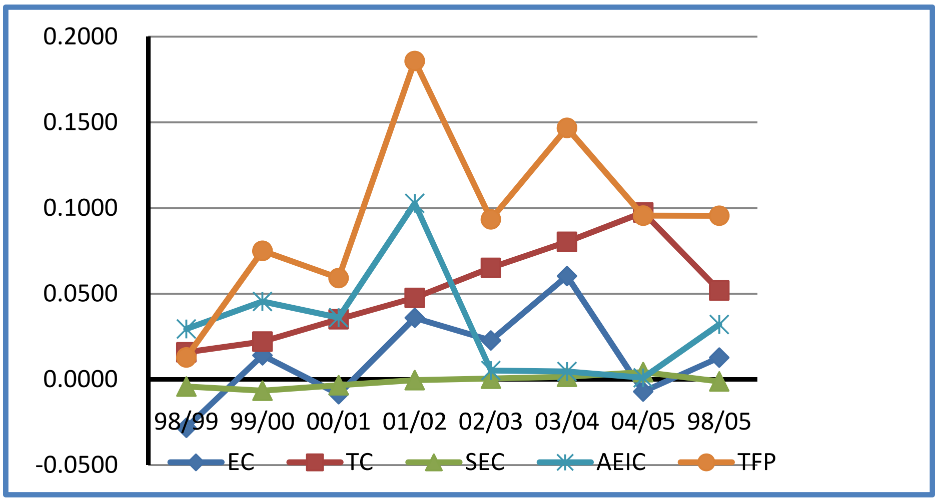

Figure 6.13: TFP change and its decomposition in Chinese banking sector

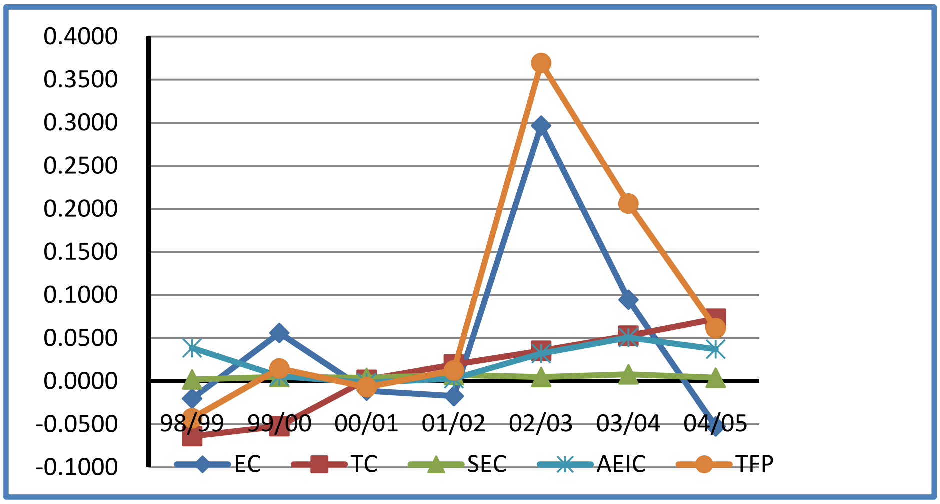

Figure 6.14: TFP change and its decomposition in Hong Kong's banking sector

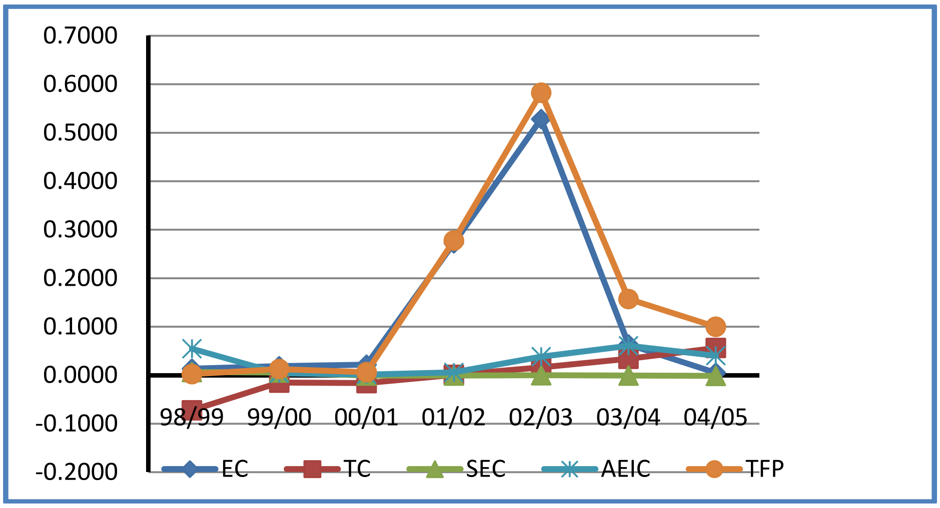

Figure 6.15: TFP change and its decomposition in Indian banking sector

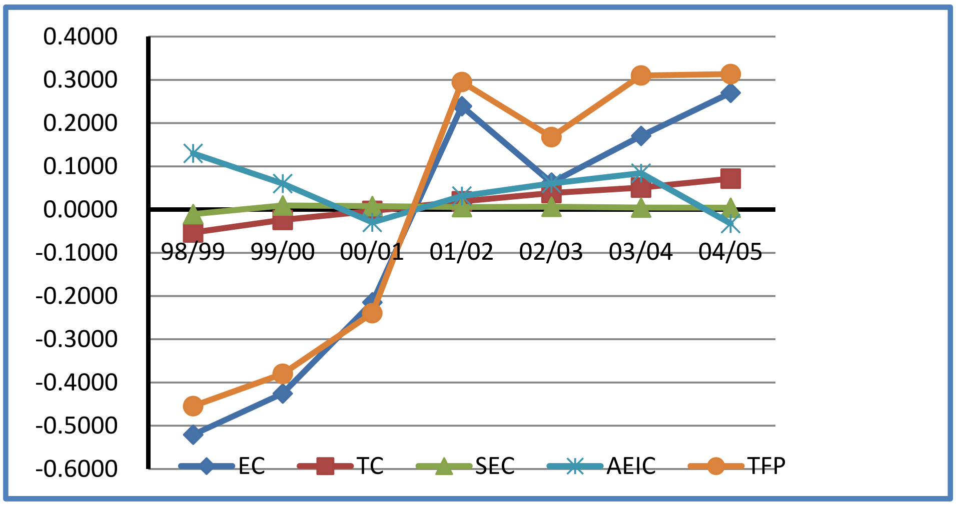

Figure 6.16: TFP change and its decomposition in Indonesian banking sector

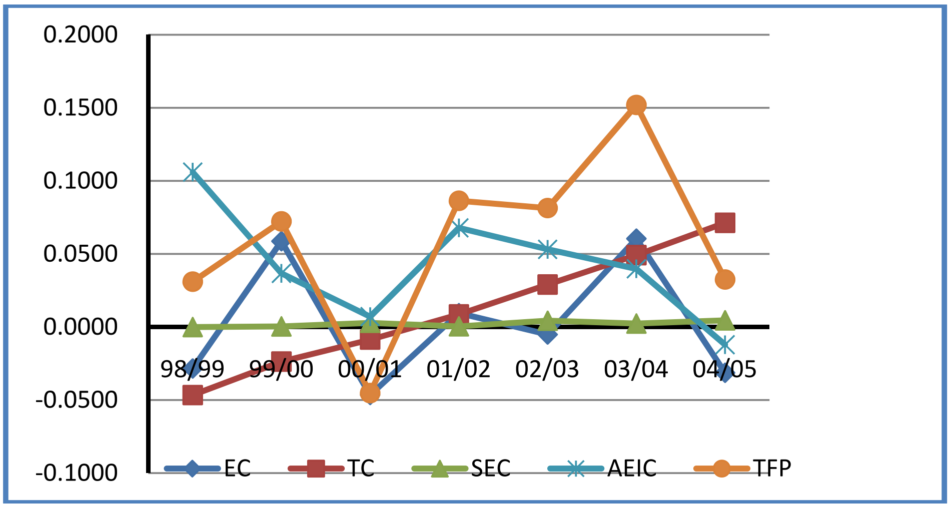

Figure 6.17: TFP change and its decomposition in Malaysian banking sector

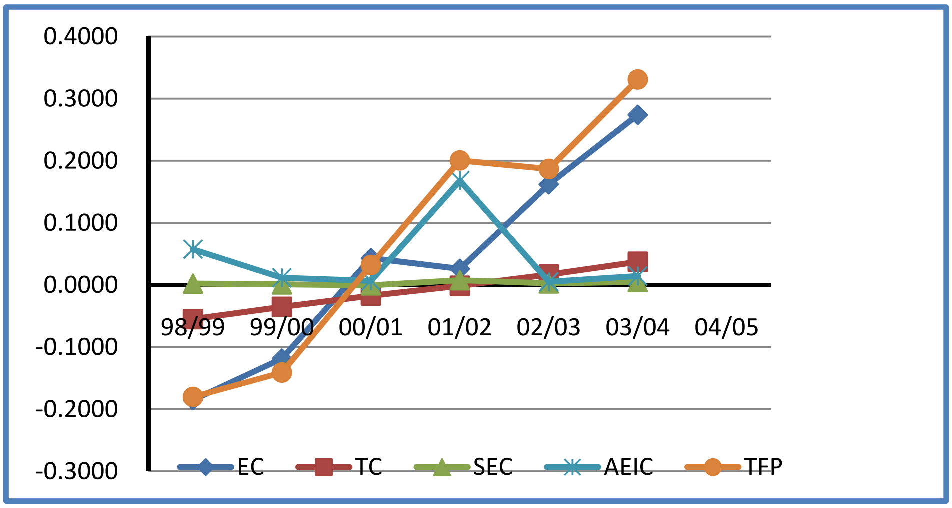

Figure 6.18: TFP change and its decomposition in Philippine banking sector

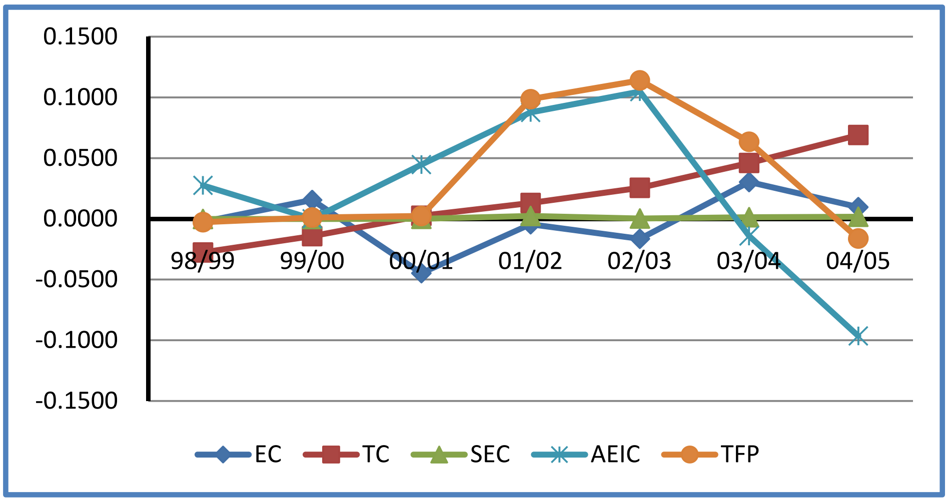

Figure 6.19: TFP change and its decomposition in Singaporean banking sector

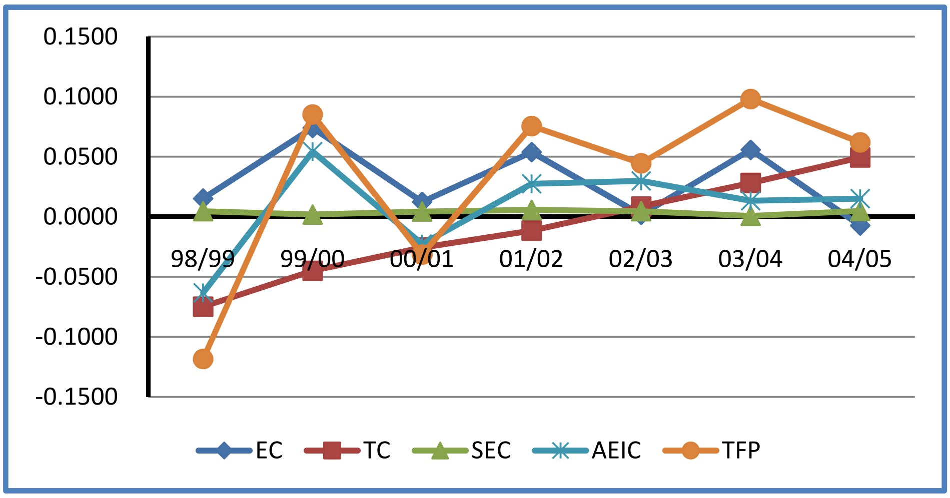

Figure 6.20: TFP change and its decomposition in Korean banking sector

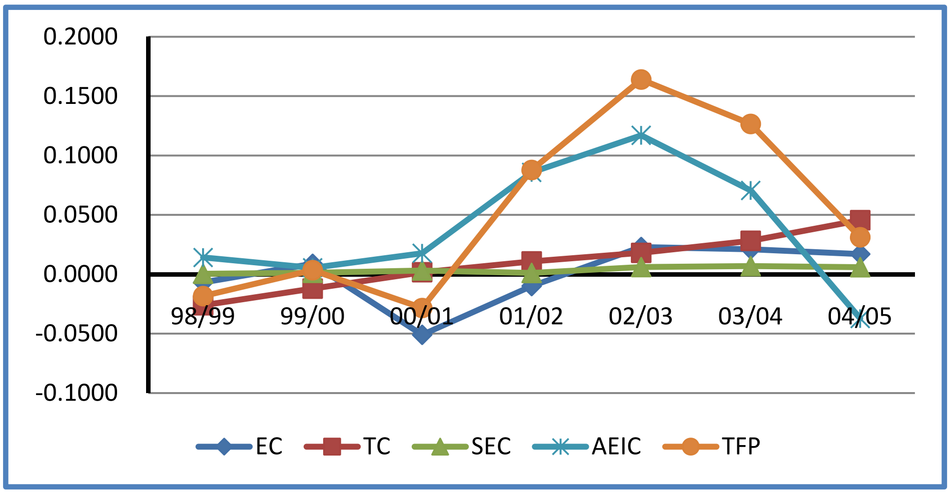

Figure 6.21: TFP change and its decomposition in Taiwanese banking sector

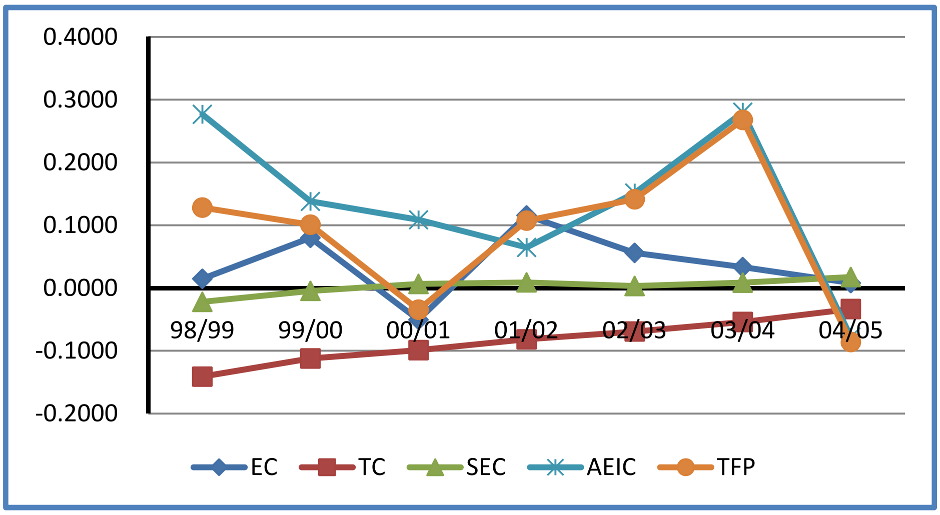

Figure 6.22: TFP change and its decomposition in Thai banking sector

have experienced positive productivity growth over the entire sample period[21], especially for India and China as their productivity performance rank first and second compared to other Asian banking industries.

6.6 Conclusion

In this chapter, effects of exogenous influences on both inefficiency and random noise error term are discussed. Specifically, exogenous influences can be placed in three positions, either the mean of truncated distribution of inefficiency as explanatory variables to the variation of the inefficiency scores, or/and the variance of truncated distribution of inefficiency, or/and the variance of normal distribution of the two-sided error term. Distinguished from empirical study implemented in the previous chapters, in which exogenous influences are considered to affect the shape of the frontier and modeled as additional independent variable in the stochastic frontier cost model that

[21] Singapore does appear to have a negative productivity growth in 2005. But it is mainly caused by the same number of observations in 2005 (i.e. 3 Singaporean banks are observed from 1998 to 2004 but only 1 for 2005).

eventually generate the cost efficiency estimates that are net of exogenous influences, this empirical chapter attempt to estimate the stochastic frontier model with consideration of exogenous influences on both the inefficiency term and the two-sided error term, which consequently will generate the cost efficiency estimates that incorporate the effects of exogenous variables.

Results from comparisons of models that incorporate different specifications of exogenous influences on both error components suggest that the general model that nested the exogenous influences on all three positions fit the sample data that consists of 280 commercial banks in ten major Asian banking industries.

Cost efficiency and the TFP change and its decomposition are estimated and calculated from the parameter estimates from the general model. The overall cost efficiency score is about 0.5412 and changes in a non-monotonic pattern over the sample period (decreasing from 1998 to 2001 and then increasing until 2004 with a slight decline in 2005). This non-monotonic trend is highly influenced by the 1997 Asian financial crisis as we discussed above. The overall TFP change shows around 6.5% improvement over the sample period in a non-monotonic trend. Its pattern is mainly determined by the cost efficiency change and its scale is jointly affected by the effects of technological change, scale effect change and input mix allocative efficiency change. Moreover, these sources of the overall TFP change affect the country specific TFP change in a different way.

Chapter 7 Conclusive remarks

7.1 Summary of this thesis

The literature on efficiency and productivity analysis attempts to examine the policy effect on firm's operational performance, to address the policy implications and to identify the best practice (efficient) firms by measuring the extent to which a firm is performing away from the efficient frontier. Different efficiency measures have been utilized in the past fifty years thanks to the pioneering work of Farrell (1957) who first introduced the idea of economic efficiency that consists of two components, technical efficiency and allocative efficiency. A firm is said to be technical inefficient either by producing less than maximum output from a given set of inputs or using more than the minimum input required for a given level of output. Even if the firm is technical efficient, it may still not be economic efficient due to the possible allocative inefficiency from utilizing the wrong mix of input given their prices. With information on input prices and output prices available, cost efficiency and profit efficiency can also be measured. Cost efficiency measures the extent to which a bank's cost is close to the "best" performing bank for a given level of output in the same conditions. Profit efficiency measures how close a bank is to producing the maximum possible profit given a particular level of input prices and output prices (and other variables). These two efficiency measures are particularly useful when firms' manager try not only to maximize firms' production but also to minimize production cost and maximize profit.

Like other regulated industries, banking sector has experienced dramatic changes in designs of regulatory framework and operating environment through technological progress, deregulation, privatization and inter-country and cross-border mergers and acquisitions. These global phenomenons in banking industries draw interests from economists and researchers and a large volume of efficiency and productivity studies have been seen in the literature with attempts to examine effects from these regulatory and operational events and to provide further policy implications by measuring economies of scale and scope, technical (or cost, or profit) efficiency and productivity change using non-parametric or parametric frontier approaches. The pros and cons of frontier approaches have been explicitly discussed in section 2.2 of Chapter 2.

No matter what frontier approaches are being used, the existing banking efficiency and productivity literature has mainly focused on banks in the US and European Union. Despite the recent increase in the number of studies in emerging and developing banking markets, such volume is relatively small compared to that in the US and European Union as evidenced in my literature survey of banking efficiency studies in section 2.1 of Chapter 2. In particular, there have been no cross-country comparisons of banking efficiency and productivity analysis in major emerging Asian economies.

To fill this literature gap, this thesis is amongst the first to measure and compare the cost efficiency and total factor productivity change in ten Asian banking industries by using the panel data stochastic frontier approach. This sample data consist of 280 commercial banks in ten Asian countries and regions from 1998 to 2005. Following the modern banking theory that views banks as financial intermediaries, this study focuses on retail commercial banks and excludes those banks whose loan-deposit ratio is higher than one. According to financial intermediation theory, banks' outputs are loans, other earning assets, and non-interest income, while banks' inputs are deemed to be deposits, labor and fixed assets.

Coelli *et al.* (1999) argued that exogenous environmental factors may either influence the production frontier technology or inefficiency itself as determinants. Efficiency of different airlines estimated in former case is termed as net efficiency while the latter one is named as gross efficiency. Drawing inspirations from Coelli *et al.* (1999), this thesis attempts to account influences of cross-country heterogeneous factors on cost efficiency estimates and measurement of total factor productivity change and its sources in different Asian banking industries. In this thesis, cross-country exogenous influences are modeled as seen in the two cases in Coelli *et al.* (1999) but I extend the latter one by allowing cross-country environmental variables to enter the variance of inefficiency and random noise term. This extension allows environmental influences on the scale of variations of inefficiency distribution as well as variations of random noise term. It also provides a measure to control the possible heteroscedasticity problems in econometric estimation.

First, with incorporation of cross-country heterogeneities as additional variables in the model to allow different production technology for different banking sector, this thesis measures and compares cost efficiency across ten Asian banking industries by adopting panel data stochastic frontier models. A systematic comparison is carried out between time-invariant and time-varying fixed-effects and random-effects models. Results coincide with other empirical evidence that in the circumstance of unobservable time-invariant heterogeneities, traditional fixed- and random-effects models will underestimate cost efficiency while Greene's 'true' fixed- and random-effects models will overestimate cost efficiency due to overcompensation of time-invariant heterogeneities. Under preferred Battese and Coelli (1992) time-varying random-effects model, further empirical evidence is provided in the context of panel data stochastic framework, suggesting the significance of considering cross-country differences in international comparisons of efficiency as they do explain part of variations in inefficiency estimates from estimating a common frontier. Overall cost efficiency in these Asian banking industries is 0.5897 and decreasing over the sample period, although significant technical progress and slight economies of scale are found. Moreover, discussion on the appropriate choice of output and input measure suggests that intermediation approach should be used when only values of deposits are available. Discussion on the best choice of functional form suggests that although Fourier flexible functional form satisfies global approximation, it may fail to satisfy certain theoretical assumptions imposed on cost function. It may be better to solve problems of the poor approximation that results in large dispersion of banks' efficiency by introducing additional explanatory variables rather than using much complex functional form.

Based on cost efficiency and parameter estimates from Battese and Coelli (1992) model with incorporation of cross-country heterogeneous factors, total factor productivity change is calculated by using a cost based Malmquist type productivity index, an index number counterpart to Bauer (1990) total differential approach, which allows comparison of productivity change year by year. Total factor productivity change is further decomposed into cost efficiency change, technical change, scale effects change, input mix allocative efficiency change, and output mix allocative efficiency change. The empirical results show positive but not substantial productivity change for Asian banking industries during the sample period, which is mainly attributed to net effect of technical progress and positive scale effect change after offsetting the decreasing trend of efficiency deterioration. Inclusion of allocative efficiency change largely influences the whole pattern of the TFP change, indicating that allocative efficiency change is a very important source to explain the productivity growth.

Distinguished from the above two empirical chapters, which considers the environmental influences on production technology, the last chapter of this thesis examines the effects of cross-country exogenous influences on the shape and scale of distributions of composed error term. Exogenous influences could either the mean of truncated distribution of the inefficiency term as explanatory variables to the variation of the inefficiency scores, or/and the variance of truncated distribution of the inefficiency term, or/and the variance of normal distribution of the two-sided error term. Results from comparison of models that incorporate different specifications of exogenous influences on both error components suggest that the general nested model that nested the exogenous influences on all three locations is the best to fit our sample data. The overall cost efficiency is around 0.5412. It changes in a non-monotonic pattern over the sample period. The overall TFP change shows around 6.5% over the sample period also with a non-monotonic trend. Its pattern is mainly determined by the cost efficiency change and its scale is jointly affected by the effects of technological change, scale effect change and input market allocative efficiency change.

7.2 Directions for future research

First of all, modelling risk in banking performance measurement. An important feature of this thesis lies on the inclusion of equity capital ratio in the translog cost function specification as a control for the risk factor. As dicussed in section 2.1.2.2.2 and argued by Hughes and Mester (1993), Mester (1996) and Hughes and Mester (2008), a bank's insolvency risk depends not only on the riskiness of its portfolio but also on the amount of capital it has to absorb losses. Insolvency risk affects bank's cost and profit through intensive risk management and premium the bank has to pay for uninsured debt. A bank's capital level also directly affects its cost since it can be served as an alternative funding source. Therefore, it is important to include equity capital in performance specifications. According to Hughes and Mester (2008), for most empirical studies using cash-flow (accounting) concept of cost, which include the interests paid on deposits but ignore the cost of equity, failure to include equity capital among the inputs will generate biased efficiency results. By including level of equity as quasi-fixed input in cost function, the negative derivative of cost with respect to level of equity capital will provide a measure of shadow price of equity capital. The shadow price of equity capital will equal to the market price when the amount of equity minimize costs and maximize profits. Even when these objectives cannot be conformed, the shadow price of equity nevertheless provides a measure of its opportunity cost. Hughes *et al.* (2001) addressed this issue by including the level of equity in the translog cost function. They found that the mean shadow price of equity for small banks is significantly smaller than for larger banks, implying that small banks over-utilize equity while larger banks under-utilize equity possibly due to the availability of deposit insurance and the Too-Big-To-Fail Doctrine. Recently, Kenjegalieva, Ravishjankar, Shen and Weyman-Jones (2009) provide the similar empirical evidence using the Portugal banking dataset and the Asian banking dataset used in this thesis.

Despite the above indirect measure of risk in efficiency and productivity analysis, direct measure of risk in banking performance can also be constructed if banks are viewed as portfolio optimizing agents. One of the few studies to introduce risk directly into an econometric frontier is that of Hughes *et al.* (2001). They developed a managerial utility function that depends on profit and input usage. By solving this constrained utility maximizing problem, a shadow return on equity can be derived from the profit share equations derived from the optimal solution by adjusting for gross returns on equity, $E(\pi_i / k_i)$. Then managerial performance related to risk can be examined by estimating a stochastic risk-return frontier, written as

$$E(\pi_i / k_i) = \alpha_0 + \alpha_1 \sigma_i + \alpha_2 \sigma_i^2 + v_i - u_i \qquad [7.1]$$

where σ_i is the standard error of the predicted return which is a measure of econometric prediction risk termed as a function of production plan and other explanatory variables and varies across banks. However, this model [7.1] poses a serious econometric problem that it is difficult to satisfy the essential assumption of uncorrelatedness of σ_i and inefficiency u_i.

However, despite the difficulty of constructing an econometric model to directly measure risk, it is possible to do it under non-parametric context. Inspiration is drawn from the literature of non-parametric efficiency methods of evaluating investment fund performance which proposes a list of outputs and inputs that incorporate a range of measures of portfolio return in the outputs and a range of risk (such as the variance of returns, the half-variance and so on) and operating and set-up cost in the inputs and measures the efficient performance using DEA model. To apply this model in banking performance measure including risk, besides the usual output and input specifications, one need to include variables such as profit and return on equity as outputs while incorporate the variance of returns as well as the operating and capital cost as inputs and then utilize DEA to obtain the efficiency level.

Second, in future, if the information on output prices are available, current research can be extended to measuring profit efficiency that not only considering the cost minimization but also the profit maximization. Since the ultimate objective of the shareholders is to maximize the profit, measuring profit efficiency appears to be more attractive as even if a bank is cost inefficient, it can be more profit efficient than those banks those are relative more cost efficient driven by the more sales revenue. Alternatively, even a bank is cost efficient, it can still be profit inefficient if it produce the wrong output mix according to output prices. With the impact of the recent financial turmoil, it could be expected that in the next few years, if banks would like to be more profit efficient, banks' executives should focus more on cutting cost since it would be very difficult to generate more revenues as before the crisis. Therefore, it would be more interesting if in future study, one could compare the cost efficiency and profit efficiency. Such comparisons could provide more valuable information in banks' management. If a bank is found to be more cost efficient as well as more profit efficient, this may imply that banks are benefiting from cutting more costs. If a bank is found to be more cost efficient in the next few years but less profit efficient, one can suggest that this bank is experiencing problems in creating good loans. If a bank is found to be less cost efficient as well as less profit efficient, one could argue that this bank is highly badly influence by the financial turmoil which may leave the bank with large amount non-performing loans. Moreover, with the available output price information, the index of total factor productivity change can be extended to incorporate the influence of output mix allocative inefficiency change.

Bibliography

Ahmad, M. and Bravo-Ureta, B. 1996, "Technical Efficiency Measures for Dairy Farms Using Panel Data: A Comparison of Alternative Model Specifications", *Journal of Productivity Analysis,* vol. 7, no. 4, pp. 399-415.

Aigner, D.J. and Chu, S.F. 1968, "On Estimating the Industry Production Function", *American Economic Review,* vol. 58, pp. 226-239.

Aigner, D., Lovell, C.A.K. and Schmidt, P. 1977, "Formulation and Estimation of Stochastic Frontier Production Function Models", *Journal of Econometrics,* vol. 6, no. 1, pp. 21-37.

Alam, I.M.S. 2001, "A Nonparametric Approach for Assessing Productivity Dynamics of Large U.S. Banks", *Journal of Money, Credit and Banking,* vol. 33, no. 1, pp. 121-39.

Allen, L. and Rai, A. 1996, "Operational Efficiency in Banking: An International Comparison", *Journal of Banking and Finance,* vol. 20, no. 4, pp. 655-672.

Altunbaş, Y. and Chakravarty, S.P. 2001, "Frontier Cost Functions and Bank Efficiency", *Economics Letters,* vol. 72, pp. 233-240.

Altunbaş, Y., Gardener, E.P.M., Molyneux, P. and Moore, B. 2001, "Efficiency in European Banking", *European Economic Review,* vol. 45, pp. 1931-1955.

Altunbaş, Y., Liu, M., Molyneux, P. and Seth, R. 2000, "Efficiency and Risk in Japanese Banking", *Journal of Banking and Finance,* vol. 24, pp. 1605-1628.

Alvarez, A., Amsler, C., Orea, L. and Schmidt, P. 2006, "Interpreting and Testing the Scaling Property in Models where Inefficiency Depends on Firm Characteristics", *Journal of Productivity Analysis,* vol. 25, no. 3, pp. 201-212.

Avkiran, N. 2000, "Rising Productivity of Australian Trading Banks under Deregulation 1986-1995", vol. 24, no. 2, pp. 122-140.

Battese, G.E. and Coelli, T.J. 1995, "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data", *Empirical Economics,* vol. 20, pp. 325-332.

Battese, G.E. and Coelli, T.J. 1992, "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India", *Journal of Productivity Analysis,* vol. 3, pp. 153-169.

Battese, G.E. and Coelli, T.J. 1988, "Prediction of Firm-level Technical Efficiencies with A Generalized Frontier Production Function and Panel Data", *Journal of Econometrics,* vol. 38, no. 3, pp. 387-399.

Battese, G.E. and Corra, G.S. 1977, "Estimation of A Production Frontier Model with Application to the Pastoral Zone of Eastern Australia", *Australian Journal of Agricultural Economics,* vol. 21, pp. 169-179.

Bauer, P.W. 1990, "Decomposing TFP Growth in the Presence of Cost Inefficiency, Nonconstant Returns to Scale, and Technological Progress", *Journal of Productivity Analysis,* vol. 1, no. 4, pp. 287-299.

Bauer, P.W., Berger, A.N. and Humphrey, D.B. 1993, "Efficiency and Productivity Growth in U.S. Banking" in *The Measurement of Productive Efficiency*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York Oxford, pp. 386-413.

Bauer, P.W. and Hancock, D. 1993, "The Efficiency of the Federal Reserve in Providing Check Processing Services", *Journal of Banking and Finance,* vol. 17, no. 2-3, pp. 287-311.

Baumol, W.J., Panzar, J.C. and Willig, R.D. 1988, *Contestable Markets and the Theory of Industry Structure,* Harcourt Brace Jovanovich, Publishers and its subsidiary, Academic Press, San Diego; London.

Berg, S.A., Forsund, F.R. and Jansen, E.S. 1992, "Malmquist Indices of Productivity Growth during the Deregulation of Norwegian Banking, 1980-89", *Scandinavian Journal of Economics,* vol. 94, no. 0, pp. S211-28.

Berger, A.N. and DeYoung, R. 1997, "Problem Loans and Cost Efficiency in Commercial Banks", *Journal of Banking and Finance,* vol. 21, no. 6, pp. 849-870.

Berger, A.N., Hanweck, G.A. and Humphrey, D.B. 1987, "Competitive Viability in Banking: Scale, Scope and Product Mix Economies", *Journal of Monetary Economics,* vol. 20, pp. 501-520.

Berger, A.N., Hasan, I. and Zhou, M. 2009, "Bank Ownership and Efficiency in China: What Will Happen in the World's Largest Nation?", *Journal of Banking and Finance,* vol. 33, no. 1, pp. 113-130.

Berger, A.N. and Humphrey, D.B. 1997, "Efficiency of Financial Institutions: International Survey and Directions for Future Research", *European Journal of Operational Research,* vol. 98, pp. 175-212.

Berger, A.N. and Humphrey, D.B. 1992, "Measurement and Efficiency Issues in Commercial Banking", in *Output Measurement in the Service Sectors*, ed. Z. Griliches, The University of Chicago Press, Chicago, pp. 245-279.

Berger, A.N., Hunter, W.C. and Timme, S.G. 1993, "The Efficiency of Financial Institutions: A Review and Preview of Research Past, Present and Future", *Journal of Banking and Finance,* vol. 17, pp. 221-249.

Berger, A.N., Leusner, J.H. and Mingo, J.J. 1997, "The Efficiency of Bank Branches", *Journal of Monetary Economics,* vol. 40, pp. 141-162.

Berger, A.N. and Mester, L.J. 2003, "Explaining the Dramatic Changes in Performance of US Banks: Technological Change, Deregulation, and Dynamic Changes in Competition", *Journal of Financial Intermediation,* vol. 12, no. 1, pp. 57-95.

Berger, A.N. and Mester, L.J. 1999, *What Explains the Dramatic Changes in Cost and Profit Performance of the U.S. Banking Industry?*, Federal Reserve Bank of Philadelphia.

Berger, A.N. and Mester, L.J. 1997, "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?", *Journal of Banking and Finance,* vol. 21, no. 7, pp. 895-947.

Bjurek, H. 1996, "The Malmquist Total Factor Productivity Index", *Scandinavian Journal of Economics,* vol. 98, no. 2, pp. 303-13.

Bonin, J.P., Hasan, I. and Wachtel, P. 2005a, "Bank Performance, Efficiency and Ownership in Transition Countries", *Journal of Banking and Finance,* vol. 29, no. 1, pp. 31-53.

Bonin, J.P., Hasan, I. and Wachtel, P. 2005b, "Privatization Matters: Bank Efficiency in Transition Countries", *Journal of Banking and Finance,* vol. 29, no. 8-9, pp. 2155-2178.

Canhoto, A. and Dermine, J. 2003, "A Note on Banking Efficiency in Portugal, New vs. Old Banks", *Journal of Banking and Finance,* vol. 27, no. 11, pp. 2087-2098.

Carbo, S., Gardener, E.P.M. and Williams, J. 2002, "Efficiency in Banking: Empirical Evidence from the Savings Banks Sector", *Manchester School,* vol. 70, no. 2, pp. 204-228.

Carvallo, O. and Kasman, A. 2005, "Cost Efficiency in the Latin American and Caribbean Banking Systems", *Journal of International Financial Markets, Institutions and Money,* vol. 15, no. 1, pp. 55-72.

Casu, B., Girardone, C. and Molyneux, P. 2004, "Productivity Change in European Banking: A Comparison of Parametric and Non-parametric Approaches", *Journal of Banking and Finance,* vol. 28, no. 10, pp. 2521-2540.

Casu, B. and Girardone, C. 2004, "Large Banks' Efficiency in the Single European Market", *Service Industries Journal,* vol. 24, no. 6, pp. 129-142.

Casu, B. and Girardone, C. 2002, "A Comparative Study of the Cost Efficiency of Italian Bank Conglomerates", *Managerial Finance,* vol. 28, pp. 3-23(21).

Caudill, S.B. and Ford, J.M. 1993, "Biases in Frontier Estimation due to Heteroscedasticity", *Economics Letters,* vol. 41, no. 1, pp. 17-20.

Caudill, S.B., Ford, J.M. and Gropper, D.M. 1995, "Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroscedasticity", *Journal of Business and Economic Statistics,* vol. 13, no. 1, pp. 105-111.

Cavallo, L. and Rossi, S.P.S. 2002, "Do Environmental Variables Affect the Performance and Technical Efficiency of the European Banking Systems? A Parametric Analysis Using the Stochastic Frontier Approach", *The European Journal of Finance,* vol. 8, no. 1, pp. 123.

Caves, D.W., Christensen, L.R. and Diewert, W.E. 1982, "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity", *Econometrica,* vol. 50, no. 6, pp. 1393-1414.

Chang, C.E., Hasan, I. and Hunter, W.C. 1998, "Efficiency of Multinational Banks: An Empirical Investigation", *Applied Financial Economics,* vol. 8, pp. 689-696.

Charnes, A., Cooper, W.W. and Rhodes, E. 1979, "Measuring the Efficiency of Decision-Making Units", *European Journal of Operational Research,* vol. 3, no. 4, pp. 339-339.

Christopoulos, D.K., Lolos, S.E.G. and Tsionas, E.G. 2002, "Efficiency of the Greek Banking System in View of the EMU: A Heteroscedastic Stochastic Frontier Approach", *Journal of Policy Modeling,* vol. 24, no. 9, pp. 813-829.

Christopoulos, D.K. and Tsionas, E.G. 2001, "Banking Economic Efficiency in the Deregulation Period: Results from Heteroscedastic Stochastic Frontier Models", *Manchester School,* vol. 69, no. 6, pp. 656-676.

Clark, J.A. 1996, "Economic Cost, Scale Efficiency, and Competitive Viability in Banking", *Journal of Money, Credit and Banking,* vol. 28, no. 3, pp. 342-364.

Clark, J.A. 1988, "Economies of Scale and Scope at Depository Financial Institutions: A Review of the Literature", *Economic Review- Federal Reserve Bank of Kansas City,* vol. 73, no. 8, pp. 16-33.

Coelli, T.J., Estache, A., Perelman, S. and Trujillo, L. 2003, *A Primer on Efficiency Measurement for Utilities and Transport Regulators,* The World Bank, Washington, D.C.

Coelli, T.J., Perelman, S. and Romano, E. 1999, "Accounting for Environmental Influences in Stochastic Frontier Models: With Application to International Airlines", *Journal of Productivity Analysis,* vol. 11, no. 3, pp. 251-273.

Coelli, T.J., Rao, D.S.P., O'Donnell, C.J. and Battese, G.E. 2005, *An Introduction to Efficiency and Productivity Analysis,* Second Edition, Springer, USA.

Cornwell, C., Schmidt, P. and Sickles, R.C. 1990, "Production Frontiers with Cross-Sectional and Time-Series Variations in Efficiency Levels", *Journal of the Royal Statistical Society,* vol. 46, pp. 185-200.

Cuesta, R.A. and Orea, L. 2002, "Mergers and Technical Efficiency in Spanish Savings Banks: A Stochastic Distance Function Approach", *Journal of Banking and Finance,* vol. 26, no. 12, pp. 2231-2247.

Debreu, G. 1951, "The Coefficient of Resource Utilization", *Econometrica,* vol. 19, no. 3, pp. 273-292.

Demsetz, H. 1973, “Industry Structure, Market Rivalry, and Public Policy”, *Journal of Law and Economics*, vol. 16, pp. 1-9.

Denny, M., Fuss, M. and Waverman, L. 1981, "The Measurement of Total Factor Productivity in Regulated Industries, with An Application to Canadian Telecommunications" in *Productivity Measurement in Regulated Industries*, eds. T.G. Cowing and R.E. Stevenson, Academic Press, New York.

Deprins, D. and Simar, L. 1989a, "Estimating Technical Inefficiencies with Corrections for Environmental Conditions with An Application to Railway Companies", *Annals of Public and Cooperative Economics,* vol. 60, no. 1, pp. 81-102.

Deprins, D. and Simar, L. 1989b, "Estimation de frontières déterministes avec facteurs exogénes d'inefficacité", *Annales d'Economie et de Statistique,* , no. 14.

Devaney, M. and Weber, W.L. 2000, "Productivity Growth, Market Structure, and Technological Change: Evidence from the Rural Banking Sector", *Applied Financial Economics,* vol. 10, no. 6, pp. 587-95.

Dietsch, M. and Lozano-Vivas, A. 2000, "How the Environment Determines Banking Efficiency: A Comparison between French and Spanish Industries", *Journal of Banking and Finance,* vol. 24, no. 6, pp. 985-1004.

Diewert, W.E. 1989, *The Measurement of Productivity*, UBC Department of Economics.

Diewert W.E. 1976, "Exact and Superlative Index Numbers", *Journal of Econometrics*, Vol. 4 pp.114 - 145.

Färe, R. and Grosskopf, S. 1996, *Intertemporal Production Frontiers: with Dynamic DEA,* Kluwer Academic Publishers, Boston.

Färe, R., Grosskopf, S., Lindgren, B. and Roos, P. 1994a, "Productivity Developments in Swedish Hospitals: A Malmquist Output Index Approach" in *Data Envelopment Analysis: Theory, Methodology and Application*, eds. A. Charnes, W.W. Cooper, A.Y. Lewin and L.M. Seiford, Kluwer Academic Publishers, Boston.

Färe, R., Grosskopf, S. and Norris, M. 1997, "Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries: Reply", *The American Economic Review,* vol. 87, no. 5, pp. 1040-1044.

Färe, R., Grosskopf, S., Norris, M. and Zhang, Z. 1994b, "Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries", *The American Economic Review,* vol. 84, no. 1, pp. 66-83.

Farrell, M.J. 1957, "The Measurement of Productive Efficiency", *Journal of the Royal Statistical Society,* vol. 120, pp. 253-281.

Fecher, F. and Perelman, S. 1992, "Productivity Growth and Technical Efficiency in OECD Industrial Activities" in *Industrial Efficiency in Six Nations*, ed. R.E. Caves, MIT Press, Cambridge and London, pp. 459.

Fecher, F. and Pestieau, P. 1993, "Efficiency and Competition in O.E.C.D. Financial Services" in *The Measurement of Productive Efficiency: Techniques and Applications*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York Oxford, pp. 375-385.

Ferrier, G.D. and Lovell, C.A.K. 1990, "Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence", *Journal of Econometrics,* vol. 46, no. 1-2, pp. 229-245.

Førsund, F.R. and Hjalmarsson, L. 1979, "Frontier Production Functions and Technical Progress: A Study of General Milk Processing in Swedish Dairy Plants", *Econometrica,* vol. 47, no. 4, pp. 883-900.

Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. 2008, "Efficiency and Productivity" in *The Measurement of Productive Efficiency and Productivity Growth*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York, pp. 3-91.

Friedman, M. 1956 "The Quantity Theory of Money-A Restatement," in Studies in the Quantity Theory of Money, ed., Milton Friedman, University of Chicago Press, Chicago, 1956, pp. 3-21.

Fries, S. and Taci, A. 2005, "Cost Efficiency of Banks in Transition: Evidence from 289 Banks in 15 Post-communist Countries", *Journal of Banking and Finance,* vol. 29, no. 1 SPEC. ISS., pp. 55-81.

Fu, X. and Heffernan, S. 2007, "Cost X-Efficiency in China's Banking Sector", *China Economic Review,* vol. 18, no. 1, pp. 35-53.

Fukuyama, H. 1995, "Measuring Efficiency and Productivity Growth in Japanese Banking: A Nonparametrie Frontier Approach", *Applied Financial Economics,* vol. 5, no. 2, pp. 95.

Fukuyama, H. and Weber, W.L. 2002, "Estimating Output Allocative Efficiency and Productivity Change: Application to Japanese Banks", *European Journal of Operational Research,* vol. 137, no. 1, pp. 177-190.

Gallant, A.R. and Souza, G. 1991, "On the Asymptotic Normality of Fourier Flexible Form Estimates", *Journal of Econometrics,* vol. 50, no. 3, pp. 329-353.

Gallant, A.R. 1982, "Unbiased Determination of Production Technologies", *Journal of Econometrics,* vol. 20, pp. 285-323.

Gallant, A.R. 1981, "On the Bias in Flexible Functional Forms and An Essentially Unbiased Form", *Journal of Econometrics,* vol. 15, pp. 211-245.

Gathon, H. and Perelman, S. 1992, "Measuring Technical Efficiency in European Railways: A Panel Data Approach", *Journal of Productivity Analysis,* vol. 3, no. 1, pp. 135-151.

Gilbert, R.A. and Wilson, P.W. 1998, "Effects of Deregulation on the Productivity of Korean Banks", *Journal of Economics and Business,* vol. 50, no. 2, pp. 133-155.

Gilligan, T., Smirlock, M. and Marshall, W. 1984, "Scale and Scope Economies in the Multi-Product Banking Firm", *Journal of Monetary Economics,* vol. 13, pp. 393-405.

Girardone, C., Molyneux, P. and Gardener, E.P.M. 2004, "Analysing the Determinants of Bank Efficiency: The Case of Italian Banks", *Applied Economics,* vol. 36, no. 3, pp. 215-227.

Gong, B. and Sickles, R.C. 1992, "Finite Sample Evidence on the Performance of Stochastic Frontiers and Data Envelopment Analysis Using Panel Data", *Journal of Econometrics,* vol. 51, no. 1-2, pp. 259-284.

Greene, W.H. 2008, "The Econometric Approach to Efficiency Analysis" in *The Measurement of Productive Efficiency and Productivity Growth*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York, pp. 92-250.

Greene, W.H. 2005, "Fixed and Random Effects in Stochastic Frontier Models", *Journal of Productivity Analysis,* vol. 23, pp. 7-23.

Greene, W.H. 2004, "Distinguishing between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems", *Health Economics,* vol. 13, no. 10, pp. 959-980.

Greene, W.H. 1993, "The Econometric Approach to Efficiency Analysis" in *The Measurement of Production Efficiency: Techniques and Applications*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York Oxford, pp. 68-119.

Greene, W.H. 1990, "A Gamma-distributed Stochastic Frontier Model", *Journal of Econometrics,* vol. 46, no. 1-2, pp. 141-163.

Grifell-Tatjé, E. and Lovell, C.A.K. 1997, "The Sources of Productivity Change in Spanish Banking", *European Journal of Operational Research,* vol. 98, no. 2, pp. 364-380.

Grifell-Tatjé, E. and Lovell, C.A.K. 1996, "Deregulation and Productivity Decline: the Case of Spanish Savings Banks", *European Economic Review,* vol. 40, no. 6, pp. 1281-1303.

Grifell-Tatjé, E. and Lovell, C.A.K. 1995, "A Note on the Malmquist Productivity Index", *Economics Letters,* vol. 47, no. 2, pp. 169-175.

Grosskopf, S. 1993, "Efficiency and Productivity" in *The Measurement of Productive Efficiency: Techniques and Applications*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York Oxford, pp. 160-194.

Hadri, K. 1999, "Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function", *Journal of Business and Economic Statistics,* vol. 17, no. 3, pp. 359-363.

Hadri, K., Guermat, C. and Whittaker, J. 2003a, "Estimating Farm Efficiency in the Presence of Double Heteroscedasticity Using Panel Data", *Journal of Applied Economics,* vol. 0, pp. 255-268.

Hadri, K., Guermat, C. and Whittaker, J. 2003b, "Estimation of Technical Inefficiency Effects Using Panel Data and Doubly Heteroscedastic Stochastic Production Frontiers", *Empirical Economics,* vol. 28, no. 1, pp. 203-222.

Hahn, F.H. and Matthews, R.C.O. 1964, “The Theory of Economic Growth: A Survey”, *The Economic Journal*, vol. 74, no. 296, pp. 779-902

Hannan, T.H. and Hanweck, G.A. 1988, "Bank Insolvency Risk and the Market for Large Certificates of Deposit", *Journal of Money, Credit and Banking,* vol. 20, no. 2, pp. 203-211.

Hao, J., Hunter, W.C. and Yang, W.K. 2001, "Deregulation and Efficiency: the Case of Private Korean Banks", *Journal of Economics and Business,* vol. 53, no. 2-3, pp. 237-254.

Hasan, I. and Marton, K. 2003, "Development and Efficiency of the Banking Sector in A Transitional Economy: Hungarian Experience", *Journal of Banking and Finance,* vol. 27, no. 12, pp. 2249-2271.

Hausman, J.A. and Taylor, W.E. 1981, "Panel Data and Unobservable Individual Effects", *Econometrica,* vol. 49, no. 6, pp. 1377-1398.

Huang, T.H. 2000, "Estimating X-Efficiency in Taiwanese Banking Using A Translog Shadow Profit Function", *Journal of Productivity Analysis,* vol. 14, no. 3, pp. 225-245.

Huang, C.J. and Liu, J. 1994, "Estimation of A Non-neutral Stochastic Frontier Production Function", *Journal of Productivity Analysis,* vol. 5, no. 2, pp. 171-180.

Hughes, J.P. and Mester, L.J. 2008, *Efficiency in Banking: Theory, Practice and Evidence*, Federal Reserve Bank of Philadelphia Research Working Paper 08/1, Philadelphia, PA: Federal Reserve Bank of Philadelphia, and in *The Oxford Handbook of Banking*, Ed. A.N. Berger, P. Molyneux, and J. Wilson, Oxford, Oxford University Press, 2009 forthcoming.

Hughes, J.P. and Mester, L.J. 1998, "Bank Capitalization and Cost: Evidence of Scale Economies in Risk Management and Signaling", *Review of Economics and Statistics,* vol. 80, no. 2, pp. 314-325.

Hughes, J.P. and Mester, L.J. 1994, *Evidence on the Objectives of Bank Managers*, Working Paper of Wharton Financial Institutions Center, The Wharton School, University of Pennsylvania.

Hughes, J.P. and Mester, L.J. 1993, "A Quality and Risk-Adjusted Cost Function for Banks: Evidence on the "Too-Big-To-Fail" Doctrine", *Journal of Productivity Analysis,* vol. 4, pp. 293-315.

Hughes, J.P., Lang, W., Mester, L.J. and Choon-Geol Moon 1995, *Recovering Technologies that Account for Generalized Managerial Preferences: An Application to Non-Risk-Neutral Banks*, Working Paper, Wharton School Center for Financial Institutions, University of Pennsylvania.

Hughes, J.P., Mester, L.J. and Moon, C.G. 2001, "Are Scale Economies in Banking Elusive or Illusive: Evidence Obtained by Incorporating Capital Structure and Risk Taking into Models of Bank Production", *Journal of Banking and Finance,* vol. 25,, pp. 2169-2208.

Humphrey, D.B. 1993, "Cost and Technical Change: Effects from Bank Deregulation", vol. 4, no. 1, pp. 9-34.

Isik, I. and Kabir Hassan, M. 2003a, "Financial Deregulation and Total Factor Productivity Change: An Empirical Study of Turkish Commercial Banks", *Journal of Banking and Finance,* vol. 27, no. 8, pp. 1455-1485.

Isik, I. and Hassan, M.K. 2003b, "Financial Disruption and Bank Productivity: The 1994 Experience of Turkish Banks", *The Quarterly Review of Economics and Finance,* vol. 43, no. 2, pp. 291-320.

Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P. 1982, "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model", *Journal of Econometrics,* vol. 19, no. 2-3, pp. 233-238.

Kaparakis, E.I., Miller, S.M. and Noulas, A.G. 1994, "Short-Run Cost Inefficiency of Commercial Banks: A Flexible Stochastic Frontier Approach", *Journal of Money, Credit and Banking,* vol. 26, no. 4, pp. 875-893.

Kasman, A. and Yildirim, C. 2006, "Cost and Profit Efficiencies in Transition Banking: The Case of New EU Members", *Applied Economics,* vol. 38, no. 9, pp. 1079-1090.

Kenjegalieva, K., Ravishjankar, G., Shen, Z. and Weyman-Jones, T.G. 2009, *Modelling Risk in Efficiency and Productivity Analysis of Banking Systems*, Wolpertinger 2009 Conference Paper, European Association of Teachers of Banking and Finance University of Roma, Italy.

Kim, S., Kim, J. and Ryoo, H. 2006, "Restructuring and Reforms in the Korean Banking Industry", *BIS papers,* vol. 28.

Kim, Y. and Schmidt, P. 2000, "A Review and Empirical Comparison of Bayesian and Classical Approaches to Inference on Efficiency Levels in Stochastic Frontier Models with Panel Data", *Journal of Productivity Analysis,* vol. 14, no. 2, pp. 91-118.

Koopmans, T.C. 1951, "Analysis of Production as An Efficient Combination of Activities" in *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans, Cowles Commission for Research in Economics, Monograph no. 13, New York, pp. 33-97.

Kraft, E. and Tirtiroglu, D. 1998, "Bank Efficiency in Croatia: A Stochastic-Frontier Analysis", *Journal of Comparative Economics,* vol. 26, no. 2, pp. 282-300.

Kumbhakar, S.C. 1990, "Production Froniters, Panel Data, and Time-Varying Technical Inefficiency", *Journal of Econometrics,* vol. 46, pp. 201-211.

Kumbhakar, S.C. 1987, "The Specification of Technical and Allocative Inefficiency in Stochastic Production and Profit Frontiers", *Journal of Econometrics,* vol. 34, no. 3, pp. 335-348.

Kumbhakar, S.C. and Hjalmarsson, L. 1993, "Technical Efficiency and Technical Progress in Swedish Dairy Farms" in *The Measurement of Production Efficiency: Techniques and Applications*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York Oxford.

Kumbhakar, S.C. and Lovell, C.A.K. 2000, *Stochastic Frontier Analysis,* Cambridge University Press, Cambridge.

Kumbhakar, S.C. and Lozano-Vivas, A., 2005, "Deregulation and Productivity: The Case of Spanish Banks", *Journal of Regulatory Economics,* vol. 27, no. 3, pp. 331-351.

Kumbhakar, S.C., Ghosh, S. and McGuckin, J.T. 1991, "A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms", *Journal of Business and Economic Statistics,* vol. 9, no. 3, pp. 279-286.

Kumbhakar, S.C., Lozano-Vivas, A., Lovell, C.A.K. and Hasan, I. 2001, "The Effects of Deregulation on the Performance of Financial Institutions: The Case of Spanish Savings Banks", *Journal of Money, Credit and Banking,* vol. 33, no. 1, pp. 101-20.

Kumbhakar, S.C. and Wang, D. 2007, "Economic Reforms, Efficiency and Productivity in Chinese Banking", *Journal of Regulatory Economics,* vol. 32, no. 2, pp. 105-129.

Kwan, S.H. 2002, *The X-Efficiency of Commercial Banks in Hong Kong*, Working Paper ed., Hong Kong Institute for Monetary Research.

Kwan, S.H. and Eisenbeis, R.A. 1996, "An Analysis of Inefficiencies in Banking: A Stochastic Cost Frontier Approach", *FRSBF Economic Review,* , no. 2, pp. 16-26.

Lang, G. and Welzel, P. 1996, "Efficiency and Technical Progress in Banking Empirical Results for A Panel of German Cooperative Banks", *Journal of Banking and Finance,* vol. 20, no. 6, pp. 1003-1023.

Laureti, T. 2008, "Modelling Exogenous Variables in Human Capital Formation through A Heteroscedastic Stochastic Frontier", *International Advances in Economic Research,* vol. 14, no. 1, pp. 76-89.

Lee, Y.H. and Schmidt, P. 1993, "A Production Frontier Model with Flexible Temporal Variation in Technical Inefficiency" in *The Measurement of Production Efficiency: Techniques and Applications*, eds. H.O. Fried, C.A.K. Lovell and S.S. Schmidt, Oxford University Press, New York Oxford.

Leibenstein, H. 1966, “Allocative efficiency vs. ‘X-efficiency’”, American Economic Review, vol. 56, pp. 293-415.

Lovell, C.A.K. 2003, "The Decomposition of Malmquist Productivity Indexes", *Journal of Productivity Analysis,* vol. 20, no. 3, pp. 437-458.

Lozano-Vivas, A. 1997, "Profit Efficiency for Spanish Savings Banks", *European Journal of Operational Research,* vol. 98, no. 2, pp. 381-394.

Lozano-Vivas, A., Pastor, J.T. and Pastor, J.M. 2002, "An Efficiency Comparison of European Banking Systems Operating under Different Environmental Conditions", *Journal of Productivity Analysis,* vol. 18, no. 1, pp. 59-77.

Maggi, B. and Rossi, S.P.S. April 2003, *An Efficiency Analysis of Banking System: A Comparison of European and United States Large Commercial Banks Using Different Functional Forms*, Working Paper edn, Department of Economics, University of Vienna.

Malmquist, S. 1953, "Index Numbers and Indifference Surfaces", *Trabajos de Estatistica,* vol. 4, no. 2, pp. 209-242.

Maudos, J., Pastor, J.M., Pérez, F. and Quesada, J. 2002, "Cost and Profit Efficiency in European Banks", *Journal of International Financial Markets, Institutions and Money,* vol. 12, pp. 33-58.

Maudos, J. and Pastor, J.M. 2001, "Cost and Profit Efficiency in Banking: An International Comparison of Europe, Japan and the USA", *Applied Economics Letters,* vol. 8, no. 6, pp. 383-387.

McAllister, P.H. and McManus, D. 1993, "Resolving the Scale Efficiency Puzzle in Banking", *Journal of Banking and Finance,* vol. 17, pp. 389-405.

McFadden, D. 1978, "Cost Revenue and Profit Functions" in *Production Economics: a dual approach to theory and applications*, eds. M. Fuss and D. McFadden,Amsterdam, North Holland.

Meeusen, W. and van Den Broeck, J. 1977, "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error", *International Economic Review,* vol. 18, no. 2, pp. 435-444.

Mertens, A. and Urga, G. 2001, "Efficiency, Scale and Scope Economies in the Ukrainian Banking Sector in 1998", *Emerging Markets Review,* vol. 2, no. 3, pp. 292-308.

Mester, L.J. 1996, "A Study of Bank Efficiency Taking into Account Risk-Preferences", *Journal of Banking and Finance,* vol. 20, no. 6, pp. 1025-1045.

Mitchell, K. and Onvural, N.M. 1996, "Economies of Scale and Scope at Large Commercial Banks: Evidence from the Fourier Flexible Functional Form", *Journal of Money, Credit and Banking,* vol. 28, no. 2, pp. 178-199.

Molyneux, P., Altunbaş, Y. and Gardener, E. 1996, *Efficiency in European Banking,* John Wiley and Sons, Chichester.

Mukherjee, K., Ray, S.C. and Miller, S.M. 2001, "Productivity Growth in Large US Commercial Banks: The Initial Post-deregulation Experience", *Journal of Banking and Finance,* vol. 25, no. 5, pp. 913-939.

Murillo-Melchor, C., Pastor, J.M. and Tortosa-Ausina, E. 2005, "Productivity Growth in European Banking", *Working paper,* .

Nishimizu, M. and Page, J.M.,Jr. 1982, "Total Factor Productivity Growth, Technological Progress and Technical Efficiency Change: Dimensions of Productivity Change in Yugoslavia, 1965-78", *The Economic Journal,* vol. 92, no. 368, pp. 920-936.

Noulas, A.G. 1997, "Productivity Growth in the Hellenic Banking Industry: State versus Private Banks", *Applied Financial Economics,* vol. 7, no. 3, pp. 223.

Orea, L. 2002, "Parametric Decomposition of A Generalized Malmquist Productivity Index", *Journal of Productivity Analysis,* vol. 18, no. 1, pp. 5-22.

Pastor, J., Pérez, F. and Quesada, J. 1997, "Efficiency Analysis in Banking Firms: An International Comparison", *European Journal of Operational Research,* vol. 98, no. 2, pp. 395-407.

Pastor, J.M. and Serrano, L. 2005, "Efficiency, Endogenous and Exogenous Credit Risk in the Banking Systems of the Euro Area", *Applied Financial Economics,* vol. 15, no. 9, pp. 631-649.

Perera, S., Skully, M. and Wickramanayake, J. 2007, "Cost Efficiency in South Asian Banking: The Impact of Bank Size, State Ownership and Stock Exchange Listings", *International Review of Finance,* vol. 7, no. 1-2, pp. 35-60.

Perelman, S. and Pestieau, P. 1988, "Technical Performance in Public Enterprises : A Comparative Study of Railways and Postal Services", *European Economic Review,* vol. 32, no. 2-3, pp. 432-441.

Pitt, M.M. and Lee, L. 1981, "The Measurement and Sources of Technical Efficiency in the Indonesian Weaving Industry", *Journal of Development Economics,* vol. 9, pp. 43-64.

Ray, S.C. and Desli, E. 1997, "Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries: Comment", *The American Economic Review,* vol. 87, no. 5, pp. 1033-1039.

Reifschneider, D. and Stevenson, R.E. 1991, "Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency", *International Economic Review,* vol. 32, no. 3, pp. 715-723.

Resti, A. 1997, "Evaluating the Cost-Efficiency of the Italian Banking System: What can be Learned From the Joint Application of Parametric and Non-Parametric Techniques", *Journal of Banking and Finance,* vol. 21, no. 2, pp. 221-250.

Rungsuriyawiloon, S. and Coelli, T. 2007, "Regulatory Reform and Economic Performance in US Electricity Generation" in *Performance Measurement and Regulation of Network Utilities*, eds. T. Coelli and D. Lawrence, Edward Elgar Publishing, Incorporated, UK, pp. 267-296.

Saal, D.S. and Parker, D., 2007, "Determining the Contribution of Technical Change, Efficiency Change and Scale Change to Productivity Growth in the Privatized English and Welsh Water and Sewerage Industry: 1985-2000", *Journal of Productivity Analysis*, vol. 28, pp. 127-139.

Schmidt, P. and Sickles, R.C. 1984, "Production Frontiers and Panel Data", *Journal of Business and Economics Statistics,* vol. 2, no. 4, pp. 367-374.

Stiroh, K.J. 2000, "How Did Bank Holding Companies Prosper in the 1990s?", *Journal of Banking and Finance,* vol. 24, no. 11, pp. 1703-1745.

Sealey, C.W. and Lindley, J.T. 1977, "Inputs, Outputs, and A Theory of Production and Cost at Depository Financial Institutions", *The Journal of Finance,* vol. 32, no. 4, pp. 1251-1266.

Sensarma, R. 2006, "Are Foreign Banks always the Best? Comparison of State-Owned, Private and Foreign Banks in India", *Economic Modelling,* vol. 23, no. 4, pp. 717-735.

Shen, Z., Liao, H. and Weyman-Jones, T. 2009, "Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions", *Journal of Chinese Economic and Business Studies*, vol. 7, no. 2, pp 199-218.

Shen, Z., Liao, H. and Weyman-Jones, T. 2009, "Cost Efficiency Analysis in Banking Industries of Ten Asian Countries and Regions" in *China's Three Decades of Economic Reforms*, eds. X.H. Liu and W. Zhang, Routledge, Taylor & Francis Group, UK.

Solow, R.A. 1957, "Technical Change and the Aggregate Production Function", *Review of Economics and Statistics,* vol. 39, pp. 312-320.

Sturm, J. and Williams, B. 2004, "Foreign Bank Entry, Deregulation and Bank Efficiency: Lessons from the Australian Experience", *Journal of Banking and Finance,* vol. 28, no. 7, pp. 1775-1799.

Tortosa-Ausina, E., Grifell-Tatje, E., Armero, C. and Conesa, D. 2008, "Sensitivity Analysis of Efficiency and Malmquist Productivity Indices: An Application to Spanish Savings Banks", *European Journal of Operational Research,* vol. 184, no. 3, pp. 1062-1084.

Tsionas, E.G., Lolos, S.E.G. and Christopoulos, D.K. 2003, "The Performance of the Greek Banking System in View of the EMU: Results from A Non-Parametric Approach", *Economic Modelling,* vol. 20, no. 3, pp. 571-592.

Vander Vennet, R. 2002, "Cost and Profit Efficiency of Financial Conglomerates and Universal Banks in Europe", *Journal of Money, Credit and Banking,* vol. 34, no. 1, pp. 254-282.

Wang, H. 2002, "Heteroscedasticity and Non-Monotonic Efficiency Effects of A Stochastic Frontier Model", *Journal of Productivity Analysis,* vol. 18, no. 3, pp. 241-253.

Wang, H. and Schmidt, P. 2002, "One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels", *Journal of Productivity Analysis,* vol. 18, no. 2, pp. 129-144.

Weill, L. 2003, "Banking Efficiency in Transition Economies: The Role of Foreign Ownership", *Economics of Transition,* vol. 11, no. 3, pp. 569-592.

Weill, L. 2004, "Measuring Cost Efficiency in European Banking: A Comparison of Frontier Techniques", *Journal of Productivity Analysis,* vol. 21, no. 2, pp. 133-152.

Wheelock, D.C. and Wilson, P.W. 1999, "Technical Progress, Inefficiency, and Productivity Change in U.S. Banking, 1984-1993", *Journal of Money, Credit and Banking,* vol. 31, no. 2, pp. 212-34.

Appendix

Appendix 1

Half-normal distribution:

$U \sim N\left(0, \sigma_u^2\right) \qquad u = |U|$
or $\quad u \sim N^+\left(0, \sigma_u^2\right)$

$$f(u) = \frac{2}{\sqrt{2\pi}\sigma_u}.\exp\left\{-\frac{u^2}{2\sigma_u^2}\right\}$$

This can be derived by:

$$f(u \mid n \geq 0) = \left[\frac{\dfrac{1}{\sqrt{2\pi}\sigma_u}.\exp\left\{-\dfrac{u^2}{2\sigma_u^2}\right\}}{\Pr(u \geq 0)}\right]$$

$$\Pr(u \geq 0) = 1 - \Pr(u < 0) = 1 - \Pr\left(\frac{u}{\sigma_u} < 0\right)$$

Where since $u \sim N\left(0, \sigma_u^2\right)$, $\dfrac{u}{\sigma_u} \sim N(0,1)$ which means that $\dfrac{u}{\sigma_u}$ follows the standard normal distribution. We know the probability and cumulative density function of standard normal distribution, that is:

Probability density function (pdf) is $\phi(z) = \dfrac{1}{\sqrt{2\pi}}.\exp\left\{-\dfrac{z^2}{2}\right\}$

Cumulative density function (cdf) is $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$

Therefore, $\Pr(u \geq 0) = 1 - \Pr\left(\dfrac{u}{\sigma} < 0\right) = 1 - \int_{-\infty}^{0} \phi\left(\dfrac{u}{\sigma}\right)dz = 1 - 0.5 = 0.5$

Then we have the probability density function for half-normal distribution, which is

$$f(u \mid n \geq 0) = \frac{2}{\sqrt{2\pi}\sigma_u}.\exp\left\{-\frac{u^2}{2\sigma_u^2}\right\}$$

and then $E(u) = \left(\dfrac{\sqrt{2}}{\sqrt{\pi}}\right)\sigma_u$, $Var(u) = \left(\dfrac{\pi - 2}{\pi}\right)\sigma_u^2$

Since we know probability density function of u, according to the definition of the mean and variance, we can derive them in a way shown followed:

$$E(u)=\int_0^\infty u\cdot f(u)du=\int_0^\infty u\cdot\frac{2}{\sqrt{2\pi}\sigma_u}\cdot e^{-\frac{u^2}{2\sigma_u^2}}=\frac{1}{\sqrt{2\pi}\sigma_u}\int_0^\infty e^{-\frac{u^2}{2\sigma_u^2}}du^2$$

$$\underset{t\in(0,\infty)}{\overset{t=u^2}{=}}\frac{1}{\sqrt{2\pi}\sigma_u}\left(-2\sigma_u^2\cdot e^{-\frac{u^2}{2\sigma_u^2}}\Big|_0^\infty\right)=\frac{1}{\sqrt{2\pi}\sigma_u}\cdot\left(2\sigma_u^2\right)=\sqrt{\frac{2}{\pi}}\sigma_u$$

$$Var(u)=\int_0^\infty (u-E(u))^2\cdot f(u)du=\int_0^\infty\left(u-\sqrt{\frac{2}{\pi}}\sigma_u\right)^2\cdot\frac{2}{\sqrt{2\pi}\sigma_u}\cdot e^{-\frac{u^2}{2\sigma_u^2}}du$$

$$=\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty\left(u^2-2\cdot\sqrt{\frac{2}{\pi}}\sigma_u\cdot u+\frac{2}{\pi}\sigma_u^2\right)\cdot e^{-\frac{u^2}{2\sigma_u^2}}du$$

$$=\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty u^2e^{-\frac{u^2}{2\sigma_u^2}}du(A)-\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty 2\sqrt{\frac{2}{\pi}}\sigma_u\cdot u\cdot e^{-\frac{u^2}{2\sigma_u^2}}du(B)+\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty\frac{2}{\pi}\sigma_u^2e^{-\frac{u^2}{2\sigma_u^2}}du(C)$$

We separate the variance term into three parts and calculate them respectively.

$$A=\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty u^2e^{-\frac{u^2}{2\sigma_u^2}}du\overset{y=u/\sigma_u}{=}\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty y^2\sigma_u^2e^{-\frac{y^2}{2}}\sigma_u dy$$

$$=\frac{2\sigma_u^2}{\sqrt{2\pi}}\int_0^\infty y^2e^{-\frac{y^2}{2}}dy=\frac{2\sigma_u^2}{\sqrt{2\pi}}(-1)\int_0^\infty ye^{-\frac{y^2}{2}}d\left(-\frac{y^2}{2}\right)=-\frac{2\sigma_u^2}{\sqrt{2\pi}}\int_0^\infty yd\left(e^{-\frac{y^2}{2}}\right)$$

Use Integration for Part

$$=-\frac{2\sigma_u^2}{\sqrt{2\pi}}\left[ye^{-\frac{y^2}{2}}\Big|_0^\infty-\int_0^\infty e^{-\frac{y^2}{2}}dy\right]=-\frac{2\sigma_u^2}{\sqrt{2\pi}}\cdot\left[0-\frac{\sqrt{2\pi}}{2}\right]=\sigma_u^2$$

$$B=\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty 2\sqrt{\frac{2}{\pi}}\sigma_u\cdot u\cdot e^{-\frac{u^2}{2\sigma_u^2}}du=2\sqrt{\frac{2}{\pi}}\sigma_u\int_0^\infty uf(u)du=2\sqrt{\frac{2}{\pi}}\sigma_u\cdot\sqrt{\frac{2}{\pi}}\sigma_u=\frac{4}{\pi}\sigma_u^2$$

$$C=\frac{2}{\sqrt{2\pi}\sigma_u}\int_0^\infty\frac{2}{\pi}\sigma_u^2e^{-\frac{u^2}{2\sigma_u^2}}du=\frac{2}{\sqrt{2\pi}\sigma_u}\cdot\frac{2}{\pi}\sigma_u^2\int_0^\infty e^{-\frac{u^2}{2\sigma_u^2}}du\overset{y=u/\sigma_u}{=}\frac{2}{\sqrt{2\pi}\sigma_u}\cdot\frac{2}{\pi}\sigma_u^2\int_0^\infty e^{-\frac{y^2}{2}}\sigma_u dy$$

$$=\frac{4}{\pi}\sigma_u^2\int_0^\infty\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy=\frac{4}{\pi}\sigma_u^2\cdot\frac{1}{2}=\frac{2}{\pi}\sigma_u^2$$

$$\textit{Therefore}\quad Var(u)=A+B+C=\frac{\pi-2}{\pi}\sigma_u^2$$

Appendix 2

$$f(\varepsilon) = \int_0^\infty f(u,\varepsilon)du$$

$$= \int_0^\infty \frac{1}{\pi\sigma_u\sigma_v} \cdot \exp\left[-\left(\frac{u^2}{2\sigma_u^2} + \frac{(\varepsilon+u)^2}{2\sigma_v^2}\right)(A)\right]du$$

Now change the form for A

$$A = \frac{u^2}{2\sigma_u^2} + \frac{(\varepsilon+u)^2}{2\sigma_v^2} = \frac{u^2}{2\sigma_u^2} + \frac{\varepsilon^2}{2\sigma_v^2} + \frac{\varepsilon u}{\sigma_v^2} + \frac{u^2}{2\sigma_v^2}$$

Let $\sigma^2 = \sigma_u^2 + \sigma_v^2$ *and then*

$$A = \frac{\sigma^2 u^2}{2\sigma_u^2\sigma_v^2} + \frac{\varepsilon u}{\sigma_v^2} + \frac{\varepsilon^2}{2\sigma_v^2}$$

$$\stackrel{\lambda=\sigma_u/\sigma_v}{=} \left(\frac{\sigma u}{\sqrt{2}\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sqrt{2}\sigma}\right)^2 - \frac{\lambda^2\varepsilon^2}{2\sigma^2} + \frac{\varepsilon^2}{2\sigma_v^2} = \left(\frac{\sigma u}{\sqrt{2}\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sqrt{2}\sigma}\right)^2 + \frac{\varepsilon^2}{2\sigma^2}$$

$$\Rightarrow f(\varepsilon) = \int_0^\infty \frac{1}{\pi\sigma_u\sigma_v} \cdot e^{-\left(\frac{\sigma u}{\sqrt{2}\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sqrt{2}\sigma}\right)^2} \cdot e^{-\frac{\varepsilon^2}{2\sigma^2}} du$$

$$= \int_0^\infty \frac{1}{\pi\sigma_u\sigma_v} \cdot e^{-\frac{\varepsilon^2}{2\sigma^2}} \cdot e^{-\left(\frac{\sigma u}{\sqrt{2}\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sqrt{2}\sigma}\right)^2} du$$

$$= \int_0^\infty \frac{1}{\pi\sigma_u\sigma_v} \cdot e^{-\frac{\varepsilon^2}{2\sigma^2}} \cdot \frac{\sigma_u\sigma_v}{\sigma} \cdot e^{-\frac{1}{2}\left(\frac{\sigma u}{\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sigma}\right)^2} d\left(\frac{\sigma u}{\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sigma}\right)$$

let $t = \left(\frac{\sigma u}{\sigma_u\sigma_v} + \frac{\lambda\varepsilon}{\sigma}\right)$, *for* $u \in (0,\infty)$, $t \in \left(\frac{\varepsilon\lambda}{\sigma}, \infty\right)$

Then $f(\varepsilon) = \int_{\frac{\varepsilon\lambda}{\sigma}}^\infty \frac{1}{\pi\sigma} \cdot e^{-\frac{\varepsilon^2}{2\sigma^2}} \cdot e^{-\frac{t^2}{2}} dt$

$$= \frac{1}{\pi\lambda} \cdot 2\pi \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\varepsilon^2}{2\sigma^2}} \cdot \int_{\frac{\varepsilon\lambda}{\sigma}}^\infty \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt$$

$$= \frac{2}{\sigma} \cdot \phi\left(\frac{\varepsilon}{\sigma}\right) \cdot \left[1 - \int_{-\infty}^{\frac{\varepsilon\lambda}{\sigma}} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt\right]$$

$$= \frac{2}{\sigma} \cdot \phi\left(\frac{\varepsilon}{\sigma}\right) \cdot \left[1 - \Phi\left(\frac{\varepsilon\lambda}{\sigma}\right)\right]$$

Appendix 3

Putting [5.15] into [5.16],

$$\ln\left(\frac{TFP(t+1)}{TFP(t)}\right)=\frac{1}{2}\sum_j\left[r_j^{t+1}+r_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)-\frac{1}{2}\sum_m\left[s_m^{t+1}+s_m^t\right]\cdot\ln\left(\frac{x_m^{t+1}}{x_m^t}\right)$$

$$=\frac{1}{2}\sum_j\left[r_j^{t+1}+r_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)-\frac{1}{2}\sum_j\left[ey_j^{t+1}+ey_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)-\frac{1}{2}\left[\frac{\partial\ln C(t+1)}{\partial t}+\frac{\partial\ln C(t)}{\partial t}\right]$$

$$+\frac{1}{2}\sum_m\left[\left(s_m^{t+1}-s_m^{*,t+1}\right)+\left(s_m^t-s_m^{*,t}\right)\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)+\ln\left(\frac{CE^{t+1}}{CE^t}\right)$$

$$=\frac{1}{2}\sum_j\left[r_j^{t+1}+r_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)-\frac{1}{2}\sum_j\left[\frac{ey_j^{t+1}}{E^{t+1}}+\frac{ey_j^t}{E^t}\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)+\frac{1}{2}\sum_j\left[\frac{ey_j^{t+1}}{E^{t+1}}+\frac{ey_j^t}{E^t}\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)$$

$$-\frac{1}{2}\sum_j\left[ey_j^{t+1}+ey_j^t\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)-\frac{1}{2}\left[\frac{\partial\ln C(t+1)}{\partial t}+\frac{\partial\ln C(t)}{\partial t}\right]+\ln\left(\frac{CE^{t+1}}{CE^t}\right)$$

$$+\frac{1}{2}\sum_m\left[\left(s_m^{t+1}-s_m^{*,t+1}\right)+\left(s_m^t-s_m^{*,t}\right)\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)$$

$$=\frac{1}{2}\sum_j\left[\frac{ey_j^{t+1}}{E^{t+1}}\left(1-E^{t+1}\right)+\frac{ey_j^t}{E^t}\left(1-E^t\right)\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)+\ln\left(\frac{CE^{t+1}}{CE^t}\right)-\frac{1}{2}\left[\frac{\partial\ln C(t+1)}{\partial t}+\frac{\partial\ln C(t)}{\partial t}\right]$$

$$+\frac{1}{2}\sum_m\left[\left(s_m^{t+1}-s_m^{*,t+1}\right)+\left(s_m^t-s_m^{*,t}\right)\right]\cdot\ln\left(\frac{w_m^{t+1}}{w_m^t}\right)+\frac{1}{2}\sum_j\left[\left(r_j^{t+1}-\frac{ey_j^{t+1}}{E^{t+1}}\right)+\left(r_j^t-\frac{ey_j^t}{E^t}\right)\right]\cdot\ln\left(\frac{y_j^{t+1}}{y_j^t}\right)$$